# A Linguistic Approach to Pitch Range Modelling

David Patterson

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

to the

University of Edinburgh

2000

# Abstract

Pitch range is currently characterised in a number of different ways across research disciplines and is often treated as a simple measurement. Pitch range has been defined as the difference between minimum and maximum f0 (Cosmides 1983). This data alone conveys no information about the distribution of f0 values within that range. Similarly the mean and standard deviation does not adequately capture important differences in the pitch range of different speakers (Ladd *et al.* 1985). Ladd (1996) describes pitch range using two partially independent dimensions of variation, that of overall level and span. This idea has been further developed by Shriberg *et al.* (1996), in a study based on a large corpus of Dutch speech. Given this two parameter model, it is possible to predict target f0 values for when speakers raise their voices from f0 values at corresponding locations in speech produced normally.

This thesis reports on three studies of pitch range variation across speakers. The experiments examine the relation between a two dimensional model of pitch range based on pitch level and pitch span with the perception of various speaker characteristics. The key to our measure of pitch range is that it is based on average data taken from clearly defined linguistic targets in speech. These targets included sentence-initial peaks, accent peaks, post-accent valleys and sentence-final lows. The results show that a pitch range model based on linguistic dimensions of variation better captures variation in listeners' judgements than the well established measures based on speakers' long term distributional properties of f0, such as 4 standard deviations around the mean, 95th-5th percentile and 90th-10th percentile.

Most importantly this thesis shows that pitch range can and should be treated as the same entity across various research disciplines - extralinguistic, paralinguistic and linguistic - rather than the current situation in which pitch range has multiple definitions depending on the particular interest of the respective research discipline.

# Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

October 2000

# Acknowledgments

Thanks first of all to my supervisor, Bob Ladd. I don't know what I have to thank him for most, the academic advice or the regular confidence counseling. I'm sure he found the latter task the most time-consuming. Thanks also to my second supervisor, Alice Turk. A good supervisorial double act which I never came to terms with and a good thing too. If ever I thought I was doing well, one would always encourage while the other made sure that I never got too cocky!

Thanks to the Carnegie Trust for the Universities of Scotland for funding me for 3 years.

Thanks to my parents and sister for all their support and help over the past few years. I may not have ever managed to completely explain what I've been doing and why I've been doing it, but the support has never wavered.

Thanks to all the members of staff and students that have helped me over the years, whether that be for comments after presentations, for doing all the hours of speech recordings, or for experimental design and statistics advice. As that accounts for everyone in the department I must offer special thanks to Eddie Dubourg for all the computing help and F1 gossip. Everyday I'd find another computing problem, and everyday Eddie found the solutions as soon as he could. Special thanks also to Ethel Jack. Without Ethel's support when things were bleak, without the regular reminders to actually do some work when I was clearly slacking too much and without being fed at lunchtimes with tasty chicken, I would never have finished my undergraduate

degree, let alone a thesis.

Thanks to all those who have passed through the Department of Linguistics at Edinburgh any time since 1993, when I started my residency in the departmental Common Room. I would especially like to thank all those that made 1994 and 1995 such a fun time that the department was impossible to leave and hence why I stayed to do a post-graduate degree. In no particular order, and from no particular year group, thanks to Dan, Dave, Miriam, Max, Diane, Catriona, Louise, Anna, Cassie, Rob, Matthew, Janet and Julie.

Since moving to the States, I'd like to thank all those that made the transition as smooth as it was. Thanks to Cindy for giving me a job, and giving me the time to finishing off my own work whenever I needed it. Thanks also to Paul, Anissa and to all those on the crew team, especially Dan, Sue and Travis.

The "extra-curricular thanks" from the Edinburgh days is centered around rowing, beer and funk. So thanks to all those that have rowed with me, drank with me and entertained me with some ugly grooves. For rowing, especially big thanks to Olly, Miles, Katrina, Magda, Katherine (well done with the silver medal at the Olympics) and the three North Americans - Jess, Eva and Garth. Even bigger thanks to Trav, Mark, Kev and Andy for helping me accomplish all I could ever want from rowing in the 1999 season. For beer, especially big thanks to Alex, Ian, Olive, Paddy, Nick, Sonas and Jen. For funk especially big thanks to Paul, Heather and Anna.

Thanks to Juliette, without whom I would never have had so much fun before, during and after my viva. Keep making me laugh, I owe you a tenner.

Not wanting to overdo the acknowledgments or anything, but thanks also to Simon who welcomed me to the department when I was just a fresher in 1991, gave me tonnes of academic assistance, a bucket load of computing help, a whole heap of life advice, an untold numbers of stubbies, multiple wah-wah contributions to the funk, and for being an awesome flatmate for 4 years. And thank goodness he never rowed!

# Contents

# List of Figures

# Chapter 1

# Introduction and Literature Review

## 1.1 Defining Pitch Range

In the most simple terms, pitch range is the difference between some topline and some bottomline from all the fundamental frequency (f0) values used by a speaker. This is as clear cut a definition of range as is currently available. There are great differences of opinion about how to define "top" and "bottom". In some sense, answering the question of how to define "top" and "bottom" is what this entire thesis is about. We will report on a comparative study to assess various suggested measures of pitch range. The strategy used to make the assessment will be to establish which measure of range most closely correlates with the perception of a selection of speaker characteristics. In doing so, we aim to take positive steps forward in the clarification of the status of pitch range in intonational phonology, psycholinguistics and speech technology.

F0 features in speech can be divided into three categories - *linguistic*, *paralinguistic* and *extralinguistic*. Linguistic features reflect the organisation of f0 into categorically distinct entities such as high tone, low tone, boundary tone and nuclear tone. Paralinguistic features mainly relate to the communcation of emotion. Extralinguistic features in speech relate to the long term voice settings which provide information

such as the sex, height, weight, health and general "character" of a person. In all these three areas of research, a notion of pitch range is often required. Because of different interests and motivations, pitch range has come to mean different things to different people. It is apparent within the literature that pitch range not only appears to be a different phenomenon to researchers between at least the linguistic research and *voice*[1] research. There is not even a consensus within each of the disciplines. We will discuss the different aspects of speech and the ensuing differences in approaches to pitch range, starting with extralinguistic, through paralinguistic, finishing with linguistic in more detail in the remainder of this chapter. Then we will show as the main body of this thesis that it may well be possible to consider a unified approach to pitch range that will be insightful to all research disciplines.

## 1.2 Pitch Range and Extralinguistic Features

### 1.2.1 Defining extralinguistic features

Laver & Hanson (1981) divide long term speaker-characterising voice features into two different sorts, namely the *"organic"* and *"phonetic"* factors. The organic features arise from anatomical differences between speakers reflecting individual differences in the geometry and dimensions of a speaker's vocal organs. These organic features set the limits of the absolute range of fundamental frequencies that a speaker is capable of producing. Therefore organic features literally set the widest definition of a topline and a bottomline for pitch range. Clearly organic features are the least informative possible measure of pitch range because they do not offer much insight into the characteristics of a speaker, or offer any insight into linguistic phenomena. The reason that organic features offer little insight is because there is a huge difference between the absolute high and low of a speaker's voice and the high and low that represents

[1] *Voice* is a single category which combines both extralinguistic and paralinguistic research when we do not need to make any distinction between the two.

the habitually used range of a speaker's voice.

The phonetic features of voice relate to the way a speaker sets his or her vocal apparatus for speaking. It immediately seems that a study of these phonetic features of voice will give a much clearer indication as to the parameters that will offer a more meaningful and practical characterisation of what pitch range might be. In trying to characterise the long term voice features of a speaker, pitch range could be measured with the topline and bottomline being the highest and lowest fundamental frequency that a speaker actually uses in speech. Here again, we stress the move from the absolute extremes of a speaker's voice to those settings habitually used in speech. This is a very simple measure of pitch range that will characterise a speaker's voice, especially for showing between-speaker differences. Phonetic features have been used regularly and to some degree successfully.

### 1.2.2 Literature Review on Extralinguistic Features

One of the clearest pieces of information communicated by pitch is the sex of a speaker. Male speakers have thicker, longer and slacker vocal folds. The range of length of vocal folds for adult males is 17-24 mm and for females it is 12.5-17 mm (Zemlin 1981). This anatomical difference is reflected in different ranges for men and women, as reported in Hollien *et al.* (1971). Hollien *et al.* measured the mean minimum and maximum fundamental frequency for a group of 332 adult males and 202 young adult females. The male range was 78-698 Hz and the female range was 139-1108 Hz. As well as the sex of a speaker, listeners' judgements of speakers' physique and age are also reasonably accurate. Laver & Trudgill (1979) cite Lass *et al.* (1978) who report that listeners typically judge weight of speakers to within 3-4 lbs (though overestimating the weight of males and underestimating the weight of females), and that they judge height of speakers to within 1.5 inches (though overestimating the height of both males and females). Further research however suggests that low f0 is incorrectly

taken to indicate large speaker body dimensions (Dommelen & Moxness 1995). Various studies (Dordain *et al.* 1967, Hollien & Shipp 1972) have shown that age is marked by pitch in both males and females. One such study investigating male speech reports that a progressive lowering of mean pitch is characteristic of male speech between the ages of 20 through to 40, then a rise in mean pitch from age 60 upwards (Hollien & Shipp 1972). This rise in mean pitch is complemented by a reduced pitch range with extreme age (Ptacek *et al.* 1966). Laver & Trudgill (1979) cite Dordain *et al.* (1967) in which they report a drop in mean pitch for older women, but a rise with extreme age. There are reports that the prediction of age from voice characteristics, including pitch range is not so clear cut. Kraayeveld (1997) reports there are conflicting results on the predictability of age due to the confounding effects of "chronological" and "physiological" age. For example, it is easy to detect whether a male's voice has broken or not (physiological), rather than to specify whether a male is exactly 11, 12, 13 or 14 years old (chronological). Kraayeveld cites the work of Ramig & Ringel (1983) who found that it is physiological aging that is more readily identified than chronological aging, as it is physiological aging that induces voice changes, as opposed to chronological aging. This view is supported by the results of Braun & Rietveld (1995) who found that it was easier to estimate the age of smokers than of non-smokers, because smokers are generally in non-optimal physical condition compared to their non-smoking counterparts. We have only made brief mention of some of the interesting results that have been related to pitch range[2].

Given the distinction between *organic* and *phonetic* factors described in section 1.2.1 we have suggested that pitch range could be characterised by two different toplines and two different bottomlines. For the topline these are the highest pitch reachable by a speaker's voice or the highest pitch reached in speech by a speaker and for the bottomline the lowest pitch reachable by a speaker's voice or the lowest pitch reached in speech by a speaker. The key point is that both the *organic* range (the more extreme measure) and the *phonetic* range (the measure related to the limits found in speech) can

[2]For a more comprehensive review see Laver & Trudgill (1979).

be considered to be characteristic of a speaker's long term setting (Laver & Trudgill 1979).

Firstly do either the *organic* or *phonetic* factors lead to an accurate measure of pitch range? A clear answer from those interested in pitch range and extralinguistic phenomenon is that the phonetic limits are a satisfactory measure of range, if all one is interested in is characterising a speaker's normal voice. Secondly, from this initial question stems the more fundamental question to this thesis; do we really know what we want a measure of pitch range to cover? Clearly limiting a speaker to just his or her "normal" voice is not characteristic of speech. In fact, in the experiments that claim to represent normal voice (Frøkjaer-Jensen & Prytz 1976, Markel *et al.* 1977), the speech used for analysis can be described as representative of only an unemotional reading voice. Unemotional speech is better described as "one emotional" speech and therefore does not represent all the characteristics of a speaker's voice. This brings into focus our first point above, which brought into question the accuracy of this measure of pitch range. We propose that an accurate understanding of pitch range would show important relationships in variation in pitch both within and across speakers. There is no mention of this in the extralinguistic research reviewed.

The second problem, being the question of how much should be incorporated into a measure of range, should now become clearer. If a speaker can be characterised as having a normal pitch range, how is this range manipulated to account for all the emotion in speech? Should we then say that a speaker has many different pitch ranges, one for each and every emotion, or should we try and define a measure of pitch range which tries to encapsulate all the possible variations which could be predicted by the emotional content being communicated?

## 1.3 Pitch Range and Paralinguistic Features

### 1.3.1 Defining paralinguistic features

Paralinguistic features of speech can be summed up in the catch phrase "it's not what you say, it's the way that you say it." Judgements as to whether a speaker is *confident, relaxed, irritated, sad* and so on all depend to a considerable extent on voice features. As cited in Allport & Cantral (1934), an author for the New York Times Magazine on June 18th, 1933 wrote, "The human voice, when the man is not making a conscious use of it by way of impersonation, does in spite of himself reflect his mood, temper and personality. It expresses the character of the man. President Roosevelt's voice reveals sincerity, good-will and kindliness, determination, conviction, strength, courage and abounding happiness." The way an utterance is spoken communicates the speaker's attitude towards himself or herself, toward the listener, and toward the message in the utterance.

If we consider extralinguistic features to be long term settings of a voice then paralinguistic features are mid term settings and concern the communication of affect by manipulations of 'tone of voice'. Although we characterise paralinguistic features as mid term settings of a speaker's voice, separating this from the long term settings of a speaker's normal voice, it is clear that in terms of actual timing, paralinguistic communication can cover a wide range of time periods. For example, one speaker can be angry in speech for a short time, maybe just loudly swearing once, while another speaker could chose to communicate his or her anger within a huge monologue. Paralinguistic features can be communicated within an utterance, by a single utterance and beyond a single utterance. If we want to describe the communication of the emotion of anger then this "is frequently conventionally conveyed, for example, by a harsh phonatory setting, with a raised pitch span and an increased loudness span" (Laver & Trudgill 1979). A key element for this thesis in Laver & Trudgill's description of the communication of anger is the "raised pitch span". We shall aim to zero in on

what exactly *span* is and how *span* can be manipulated to communicate affect through speech.

### 1.3.2 Literature review of paralinguistic features

There is a line of experimental research dating back at least to the 1930's looking into the vocal parameters of personality (Allport & Cantral 1934, Addington 1968, Brown *et al.* 1973), of accent and dialect (Lambert 1972b, Giles 1979)) and of attitudes and emotions (Fairbanks & Pronovost 1939, Davitz 1969, Bezooijen 1984, Mozziconacci 1998). Klaus Scherer has written extensively on personality markers in speech (Scherer *et al.* 1984, 1986, 1988). A review of his work and that done by others looking into vocal indicators of discrete emotions in the pre-80s is Scherer (1981). A few broad findings are agreed on, but that is by no means a suggestion that the acoustic correlates of all these features are well established. It is clear from studies (Williams & Stevens 1972, Cosmides 1983, Ladd *et al.* 1985) that a combination of prosodic features such as pitch, speech rate, rhythm and loudness contribute to the expression of emotion in speech, and that there is not one unique acoustic correlate for a particular emotion. However, what we are interested in for the purposes of this thesis are those aspects of research relating to pitch range.

Fairbanks & Pronovost (1939) recorded 6 male actors all reading the same script for each recording. The actors were asked to simulate a different emotion for each recording. The five emotions under investigation were *contempt, anger, fear, grief* and *indifference*. In the Fairbanks study two measures relating to general pitch were taken - level and range. These two measures offer greater insight into speaker characteristics than a simple range measure like the "difference between maximum f0 and minimum f0." Level captures whether a speaker has a high or low range, and range (or "span" as proposed by Ladd 1996) captures whether a speaker has a wide or a narrow range. These two measures are an important feature of pitch range studies and we discuss them in more detail in section 2.1.1. Fairbanks & Pronovost (1939) took the measure

of level to be the median f0, and the measure of range was the maximum f0 minus the minimum f0. The mean pitch level and mean total range was calculated by averaging the results for all of the six, male speakers. The results of mean total ranges, and the emotions that they are associated with are shown in table 1.1. This table shows that the average total range for all the emotions lie between 1 and 2 octaves. Fairbanks & Pronovost (1939) report their range results using whole musical tones which is slightly unusual. More recently, it has been standard practice when using a logarithmic scale to use semi-tones. For clarity, there are 12 semitones, and therefore 6 tones, in an octave. To have a range between 1 and 2 octaves therefore implies having a range between 6 and 12 tones. *Contempt, anger* (with almost identical total ranges) and *fear* have generally wider ranges than *grief* and *indifference*.

| | Emotional label | | | | |
|---|---|---|---|---|---|
| | Contempt | Anger | Fear | Grief | Indifference |
| Total Range in tones | 10.5 | 10.3 | 11.2 | 9 | 7.8 |

**Table 1.1.** Results showing mean total ranges associated with specific emotions from Fairbanks et al (1939) study

Another interesting feature of the 1939 study is that of all the three wide ranges found for *contempt, anger* and *fear*, only *contempt* was also associated with a low average pitch level. Again, we will look more closely at the interactions of level and span in section 2.1.1.

A pitch range associated with an emotion is too simple a picture. A *contempt* pitch range described as being 1.8 octaves may be true for the average range of the 6 male speakers used in the experiment, but that average range might not even be characteristic of any one of the six actors used. It is unlikely that the distance 1.8 octaves will be characteristic of every speaker's or even many speakers' pitch range for the emotion *contempt*. An average of 1.8 octaves for *contempt* tells us nothing about how *contempt* is successfully signalled by men and women with a whole host of average pitch levels and ranges. For the sake of argument, results from Fairbanks & Pronovost (1939)

could be interpreted in another way. If we take *indifference* to be the baseline neutral setting, the results in table 1.1 suggest that *grief* is communicated by a small expansion in range, *contempt* and *anger* are communicated by a larger expansion in range and *fear* by the largest expansion in range. Interpreting the results in this fashion still tells us something about the communication of emotion in speech, but takes away the specific detail like "1.8 octaves is a specific range for *contempt*", which is a result that can not be related to any actual speaker. There is another issue not considered by this early work. Even if a discrete pitch range is associated with a discrete emotion, this does not give any insight as to how the pitch range changes. Does an expansion of pitch range from 1.8 octaves to 1.9 octaves occur just by stretching the topline and bottomline, or is this expansion done in another systematic way? Diagram 1.1 reflects the various theoretically possible ways that range can be expanded.

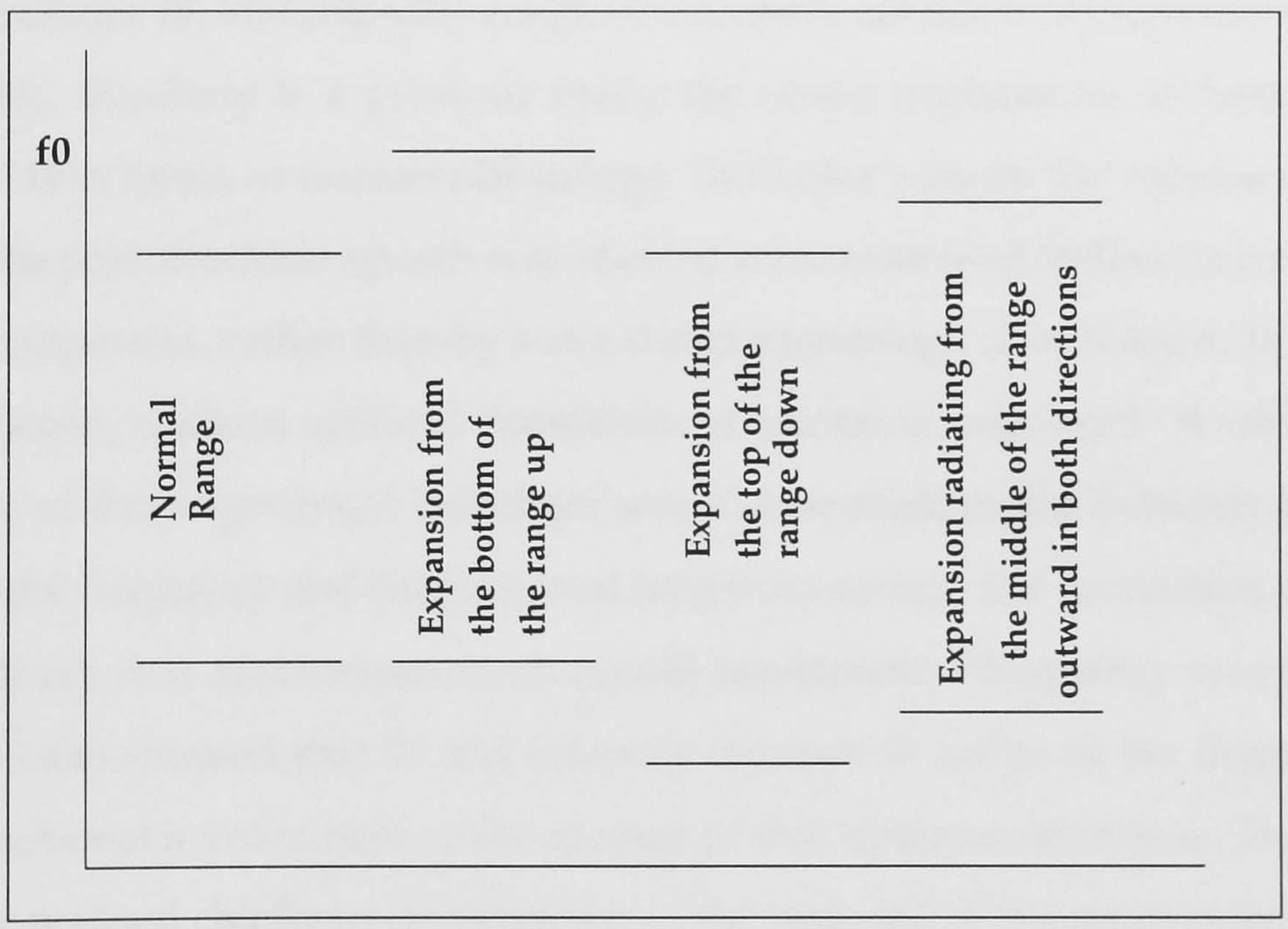



**Figure 1.1.** Theoretically possible ways in which pitch range can be expanded

The notion that the topline and bottomline of a speaker's pitch range are best characterised by the maximum f0 and the minimum f0 is not without dispute within this very early work investigating the acoustic cues to emotion in speech. Curry (1940)

says that "... the measures of the total pitch range, that is, that range which included 100 per cent of the pitches used, is not descriptive of the 'functional' pitch range... It is probable that the ranges for the median 90 per cent of the cases gives a more accurate relative expression of the differences in this respect...". It must be noted that there are two clear issues here. One relates to principled functional questions as to what pitch range actually is. The second issue relates to methodological questions such as the effective ways of getting rid of spurious outliers from a speaker's f0 distribution. The issue of what best represents the topline and bottomline of a speaker's pitch range is central to this thesis, especially as there have been differences of opinion going back to at least 1940 without a concerted effort to resolve these differences.

Huttar (1968) found that a degree of perceived emotion was found to be highly and positively correlated with f0 range (measured as the difference between the maximum f0 and minimum f0) and intensity range (the acoustic correlate of loudness, measured in decibels). Similarly to a previous study, the causal explanation of these relations suggested is in terms of human physiology. In Huttar's study the various emotional states of the person whose speech was studied were measured indirectly by means of listeners' responses, rather than by some direct physiological technique. In addition, normal speech, without artificial simulation of emotions, was used. It was noted in the results of this experiment that there was a close relationship between maximum fundamental frequency and fundamental frequency range. The correlation coefficient between these two measurements of overall fundamental frequency was 0.93. The general results showed that f0 and intensity increase in range as the degree of perceived emotional involvement of the speaker of that utterance increases. The increase in range is realised chiefly by an extension of the high end of the range rather than the low end, which is a very important point to establish in how pitch range is manipulated to convey emotion. The results, more specifically, show:

- An increase in f0 and intensity leads to an increase in happiness on a sad-happy scale.

- An increase in f0 leads to an increase in pleasure on an angry-pleased scale.
- An increase in f0 leads to an increase in boldness on an afraid-bold scale.
- An increase in f0 leads to an increase in confidence on a timid-confident scale.
- An increase in f0 leads to an increase in sureness on an unsure-sure scale.

Williams & Stevens (1972) employed two methods to investigate the effects of emotion on the acoustic characteristics of speech. They used acted speech and real-life speech. For the real-life emotional speech, Williams & Stevens used the recording of the radio announcer, describing the approach of the HINDENBURG Zeppelin, which burst into flames at Lakehurst, New Jersey on May 6th, 1937. Williams & Stevens noted that beyond the constraints of f0 changes marking principal linguistic functions, "a speaker is relatively free to use changes in f0 to convey nonlinguistic information, such as his emotions, or to convey special emphasis of some kind. Furthermore, the fundamental frequency can undergo variations that may not be intended or be under overt control of the speaker, and hence may provide an indication of the speaker's emotional state." For the measure of pitch level, they used the median f0 and for range they used the 90th percentile for the topline and the 10th percentile for the bottomline, as suggested earlier by Curry (1940). The emotions that they investigated were *anger, fear, sorrow* with a *neutral* category as well.

Their results for anger showed "the most consistent and striking acoustic manifestation of the emotion anger was a high f0 that persisted throughout the breath group. This increase was , on the average, at least half an octave above the f0 for a neutral situation. The range of the f0 observed for utterances spoken in angry situations was also considerably greater than the range for the neutral situations."

Results for fear showed that the average f0 "was lower than that observed for anger, and for some voices it was close to that for utterances spoken in neutral situations. There were, however, occasional peaks in the f0 that were much higher than those

encountered in a neutral position, These peaks were interspersed with regions where fundamental frequency was in a normal range."

For sorrow the average f0 "was considerably lower than that for neutral situations and the range of f0 was usually quite narrow."

Ladd *et al.* (1985) showed that overall range functions as a continuous variable; continuous with respect to the distinction made by Bolinger (1986) between "gradient" and "all-or-none" phenomena in intonation. The experimental procedure used to establish the continuous nature of pitch range, involved superimposing continually increasing resynthesised pitch ranges onto source utterances, and having listeners rate the utterances on a number of affect variables such as *arrogant, aroused* and *annoyed*. Ladd *et al.* (1985) report that changes in range are directly correlated with changes in the intensity of affective judgements. Further experiments with a similar methodology showed that an "annoyed, irritated, angry" voice had a higher f0 range and a harsh, pressed voice quality compared to a "normal, relaxed, friendly" voice. Range and voice quality had a strong effect on judges' inference of speaker arousal: harsh voice quality and wide range are seen as signals of arousal, annoyance and involvement. Range may be more strictly related to arousal, while voice quality has a component of positive - negative valence as well.

Leinonen *et al.* (1997) have studied the emotional variation found in the one-word utterance [saara], which is the Finnish equivalent for the name *Sarah*. They investigated the acoustic variation conveyed in ten emotional connotations that were simulated by speakers, seven women and five men. The motivation behind the choice of a one word utterance is based on the idea that the expression of emotion will be exaggerated as well as the hope of reduction of inter- and intra-subject variation. This would facilitate the identification of meaningful signal dimensions. In the Leinonen *et al.* (1997) study, the emotional connotations that were under investigation were

- *neutral, commanding, frightened, angry, astonished, scornful, sad, pleading, content,*

*admiring.*

Before analysis of the speech files to examine the acoustic correlates of emotion, a listening test, performed by 46 women and 27 men, was run, to establish that the supposed emotion had been perceived. For a good number of both male and female speakers, listeners agreed as to the emotion conveyed. All except the emotion *content* by the male speakers were agreed upon by over 50% of the listeners. Agreement was as high as 90% to 99% for some utterances. But those utterances that only 50% to 90% of listeners agreed as to the emotion conveyed were still analysed. It might be considered questionable as to whether this is an adequate level of agreement. Leinonen *et al.* (1997) does point out that Spearman's correlation coefficients suggest that both in male and female samples *commanding* was not well distinguished from *angry, content* from *admiring,* nor *pleading* from *sad* or *admiring.*

Intra-speaker comparisons of the [aa] segment showed that *commanding, angry, frightened,* and *astonished* were distinguished from *neutral* and from each other by amplitude and mean f0. *Scornful, sad, pleading, content* and *admiring* were distinguished from *neutral* by their longer duration. Specific intonation patterns were encountered for *astonished, pleading* and *scornful* connotations, with breathy or whispery phonation for *admiring.* Correlation analysis of the [aa] segment showed that mean f0 and peak volume tended to change concurrently, f0 range varied with mean f0, and variations in duration were independent of the other parameters.

The most current research on the topic of emotion in speech (Mozziconacci 1998) also shows predictable results that establish links between emotions and vocal parameters based on production and perception studies. We will just present the results for pitch range and pitch level in table 1.2, though Mozziconacci also investigated other parameters.

These results simply confirm the consistency in results associating acoustic properties of speech to the expression of emotion. With the additional information of pitch level,

| | Emotions | | | | | | |
|---|---|---|---|---|---|---|---|
| | neutrality | joy | boredom | anger | sadness | fear | indignation |
| pitch level | 65 Hz | 155 Hz | 65 Hz | 110 Hz | 102 Hz | 200 Hz | 170 Hz |
| pitch range | 5 st | 10 st | 4 st | 10 st | 7st | 8 st | 10 st |

**Table 1.2.** Results of the Mozziconacci (1998) study. The pitch level is taken to be the end of utterance low, and pitch range is measured as the difference between the maximum f0 and the minimum f0.

as shown in table 1.2, differences between emotions with similar ranges can be found. Although speakers showing the emotions of *joy, anger* and *indignation* are all shown to have pitch ranges of around 10 semitones, the speakers do vary in their pitch level, which for *joy* is 155 Hz, for *anger* is 110 Hz and for *indignation* is 170 Hz.

To sum up so far, in this introduction we have shown that studies of voice characteristics, extralinguistic and paralinguistic, often use a measure of pitch range to characterise a speaker's normal speech as a long term setting, and the expression of emotion in speech as a mid term setting. We have also shown that a measure of pitch range has gone from the widest possible definition measured as the extremes of a speaker's voice, to a narrower definition of the difference between the maximum and minimum of the fundamental frequency that a speaker habitually uses. In the majority of studies there has been no discussion of how pitch range changes between a speaker's habitual unemotional range as compared to modifications due to changing attitudes and emotions, a notable exception being Huttar (1968).

## 1.4 Pitch Range and Linguistic Features

### 1.4.1 Defining linguistic features

Liberman & Pierrehumbert (1984:157) use the term *pitch range* to refer to a "global, or at least phrase-sized, choice of pitch-scaling parameters." Other pitch features include *declination* referring to the downward trend in pitch across a phrase, *prominence* which

refers to local degree of stress or emphasis and *tune* which refers to intonation contour type. Traditionally tune is considered to fall within the scope of linguistics, while the other features are considered to be phonetic or paralinguistic. Liberman & Pierrehumbert (1984) claims that it is not possible to really understand any one of these factors without an understanding of the others as they all interact within the same system, which is the very point being made in this thesis in trying to find a suitable characterisation of pitch range to be of value for linguistics as well as voice research.

### 1.4.2 Literature review of linguistic features

Pitch range has been attributed to a general speaker characteristic and has been shown to vary due to the expression of emotion in speech. In terms of establishing the most ideal topline and bottomline, it is clear that even within an utterance the span between local maxima and minima of a speaker's f0 contour is changing. Vaissière (1983) compared the linguistic functions assigned in several languages to similar suprasegmental phenomena. Figure 1.2 (taken from Vaissière 1983) shows the general properties of f0 contours in unmarked sentences in a number of languages.

Vaissière limited her study to strictly the linguistic functions of prosody, avoiding the matter of paralinguistic functions in suprasegmental variation, although she acknowledges the importance of such variation. In her diagram she clearly shows the range of f0 variation narrows as a function of time. Although Vaissière identifies pitch range as being the difference between the final high and final low she clearly sees pitch range as being a local phenomenon that can be described as wide at the start of an utterance, becoming less so with time. There is no description of a global pitch range measure which would be of use to those interested in general voice characteristics. This thesis supports a model of pitch range which provides a global framework of pitch settings for each speaker. Within such a global framework, the variation in f0 features - the local phenomena - should be explained.

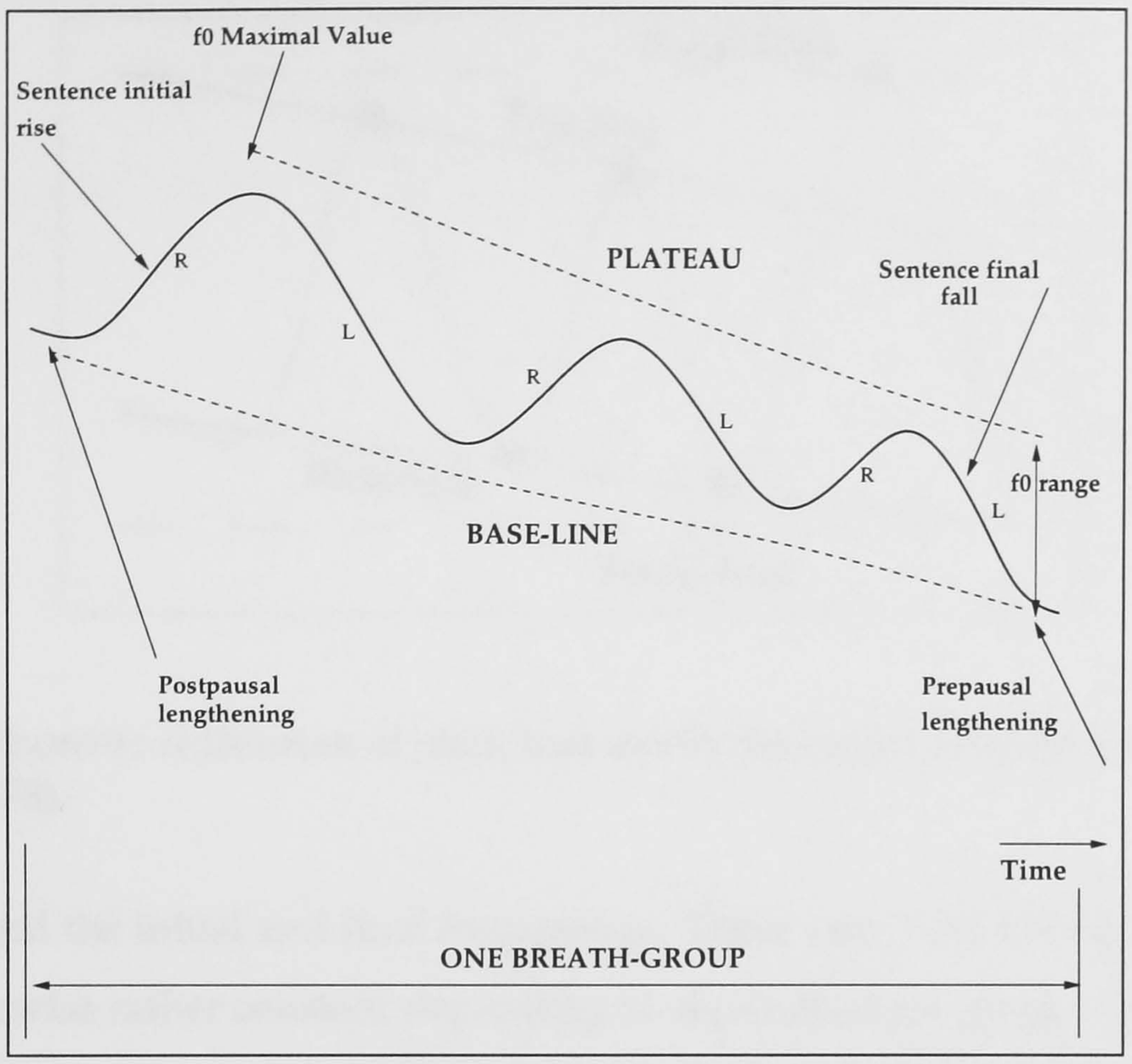



**Figure 1.2.** General Properties of f0 Contours

Bruce & Gårding (1978) set the topline and bottomline of pitch span in connection with each maximum and minimum in the f0 contour in modelling f0 for Swedish dialects. Similar settings of topline and bottomline were used in Thorsen's (1978) model of f0 for Danish. This trend of pitch range changing at such a local level led to the observation of declination (Cohen & Hart 1967). It has been shown that through the course of an utterance, even without any phonological "distractions", the topline and bottomline, delimiting local pitch movements, go down slightly. This means that a pitch movement at the beginning of a phrase will be higher than the same pitch movement later in the phrase.

In Bruce & Gårding (1978) the authors have a list of prescriptions and conventions that they use for expressing sentence intonation. "Topline and baseline are approximately straight lines. The topline connects successive f0 maxima outside the focus of a phrase. It starts and ends with the phrase. Its slope depends on the length of

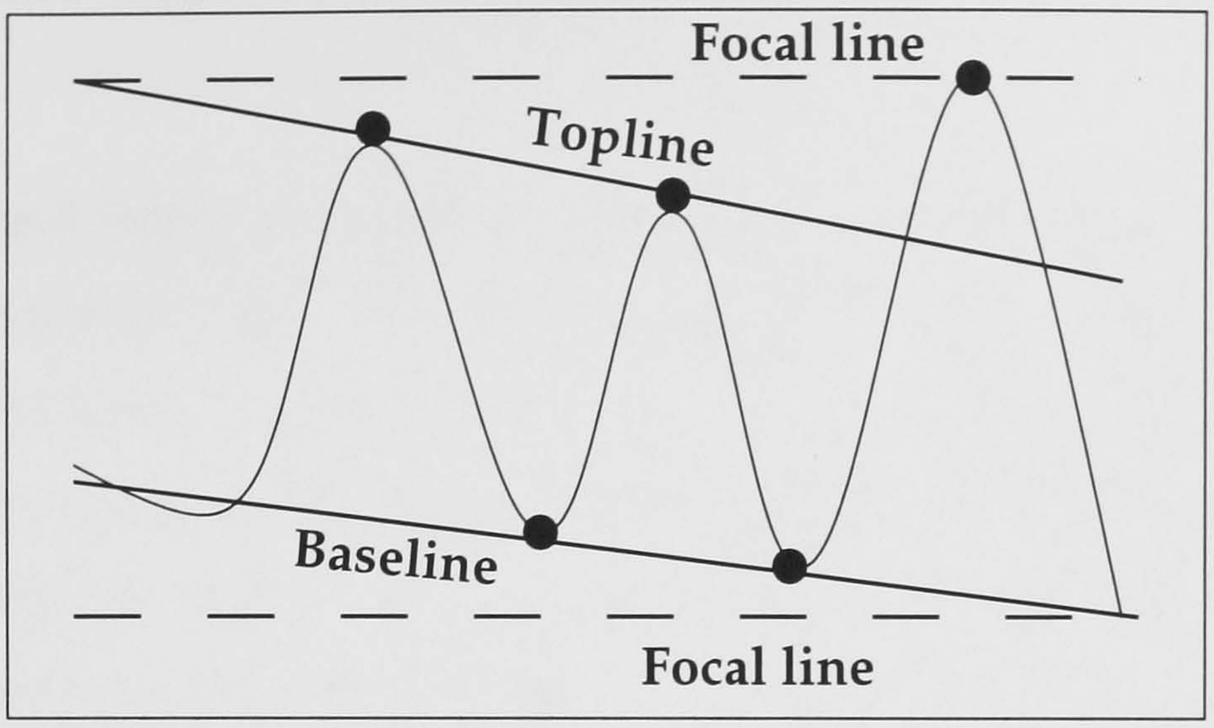



**Figure 1.3.** Phonetic realisation of pitch features in the model proposed by Bruce and Gårding, (1978).

the phrase and the initial and final frequencies. These vary with sentence intonation but are otherwise rather constant depending on the individual speaker's pitch range. The baseline connecting successive f0 minima is specified correspondingly. Focal lines connect focal maxima or minima in different sentence positions. These lines are nearly horizontal." This model is shown schematically in figure 1.3, in which the solid declining lines represent the topline and bottomline for the range within which most pitch accents are realised and the dashed lines represent the limits of the utterance range at focal maxima and minima.

There is a clear difference as to how people interpret the notion of pitch range within linguistics. Certainly in the work of Gårding (1983) and Thorsen (1978) intonational features and their mappings on to pitch contours are constrained by upper and lower limits of Gårding's tonal grid. This differs with the view of Ladd & Cutler (1983) who suggests that these grid lines are just a by-product of the linguistic specifications of accent peaks. Certainly the scaling of peaks does have intonational meaning, but also raising or expanding tonal grid lines may have more pragmatic or attitudinal effects. In the work of Gårding (1983), Thorsen (1978) and Vaissière (1983), despite slight differences of interpretation, there is reference to toplines, baselines, focus lines etc. All of these have been described as "abstractions which simplify the complexity

of actual utterance contours by indicating an overall direction or shape." (Cutler & Ladd 1983).

It has been posited that pitch range is manipulated at extremely local levels by Beckman & Pierrehumbert (1986). There has been interest in intonational phonology on the classification of two occuring features in f0 contours; one in which there is a high peak followed by a sustained high level transition to a following high peak, the other being a high peak followed by a valley then another high peak. Pierrehumbert (1980) differentiated these two distinct patterns in her typology of pitch accents, describing the first pattern as being an H*+H..H* sequence and the second one as being H*..H*. For clarity, the difference between the two accent types are drawn schematically in figure 1.4.Therefore the first accent in the first contour, with the sustained height was considered distinct from the first accent in the second contour with the post accentual valley.

This analysis was discarded by Beckman & Pierrehumbert (1986) due to theoretical problems, which are beyond the scope of this thesis. What is not beyond the scope of this thesis is the replacement analysis. Beckman & Pierrehumbert (1986) state that "we would now analyse [the sustained transition between H* accents] as involving ordinary H* accents produced in an elevated but compressed pitch range." They also state that "this reanalysis was a natural outcome of the new treatment of pitch range introduced by Liberman & Pierrehumbert (1984)."

In Liberman & Pierrehumbert (1984) a model of pitch is proposed in which f0 measurements are interpreted in terms of "a fixed baseline, a reference line that increases with pitch range, and a lowering effect specific to the domain of (certain) final pitch accents." The Liberman & Pierrehumbert (1984) model, in effect, is a model of pitch range with a topline and a bottomline (which they are calling the reference line and a baseline) which captures the characteristic of pitch range expansion which occurs from the bottom of the range upwards. Pitch range expansion will be discussed in

more detail in section 2.1.1. One of the characteristics of the Liberman & Pierrehumbert (1984) model is that both the maxima and minima of an f0 contour are measured in relation to the reference line. This characteristic differentiates it from, for example, the Bruce & Gårding (1978) model in which the topline and bottomline of pitch range was set in connection with each macimum and minimum in the f0 contour.

Returning to the issue of the Beckman & Pierrehumbert (1986) reanalysis of the H* accents; because minima are related to the reference line and not the baseline in theLiberman & Pierrehumbert model, it has to be assumed that the sustained high transition between the two H* accents is due to the f0 contour being "elevated and compressed" to the reference line. The relation between the high transition and the reference level is never explicitly stated though. It is not clear why such an f0 relation to the reference level should be attributed to pitch range. One clearly established fact which we opened this chapter with and have continued to discuss at length, is that pitch range is characterised by a topline and a bottomline. We have also assumed in this thesis that pitch range is global in nature, a view supported by Liberman & Pierrehumbert (1984). There is no indication as to how pitch range might be elevated and compressed, and this is certainly a surprise given the claim by Liberman & Pierrehumbert (1984) (cf. section 1.4.1) that made the understanding of pitch features, including pitch range, a priority. It is not clear that the local pitch range specification is justified or whether in fact the domain over which the modification discussed in Liberman & Pierrehumbert (1984) is definable.

The issue of the H*+H accent reanalysis is made clear diagramatically in figure 1.4. In figure 1.4 we have superimposed potential toplines and bottomlines to indicate possible pitch ranges. Assume that the contours represent the speech of the same speaker. The accent types in bold represent the initial Pierrehumbert (1980) analysis. Given this analysis, the pitch range for the speaker remains constant as is indicated by the postion of the topline and bottomline shown with the solid lines. This would make intuitive sense given the assumption that the speaker is the same and would have no reason to change his pitch range. On the other hand, the reanalysis of Beckman & Pierrehumbert

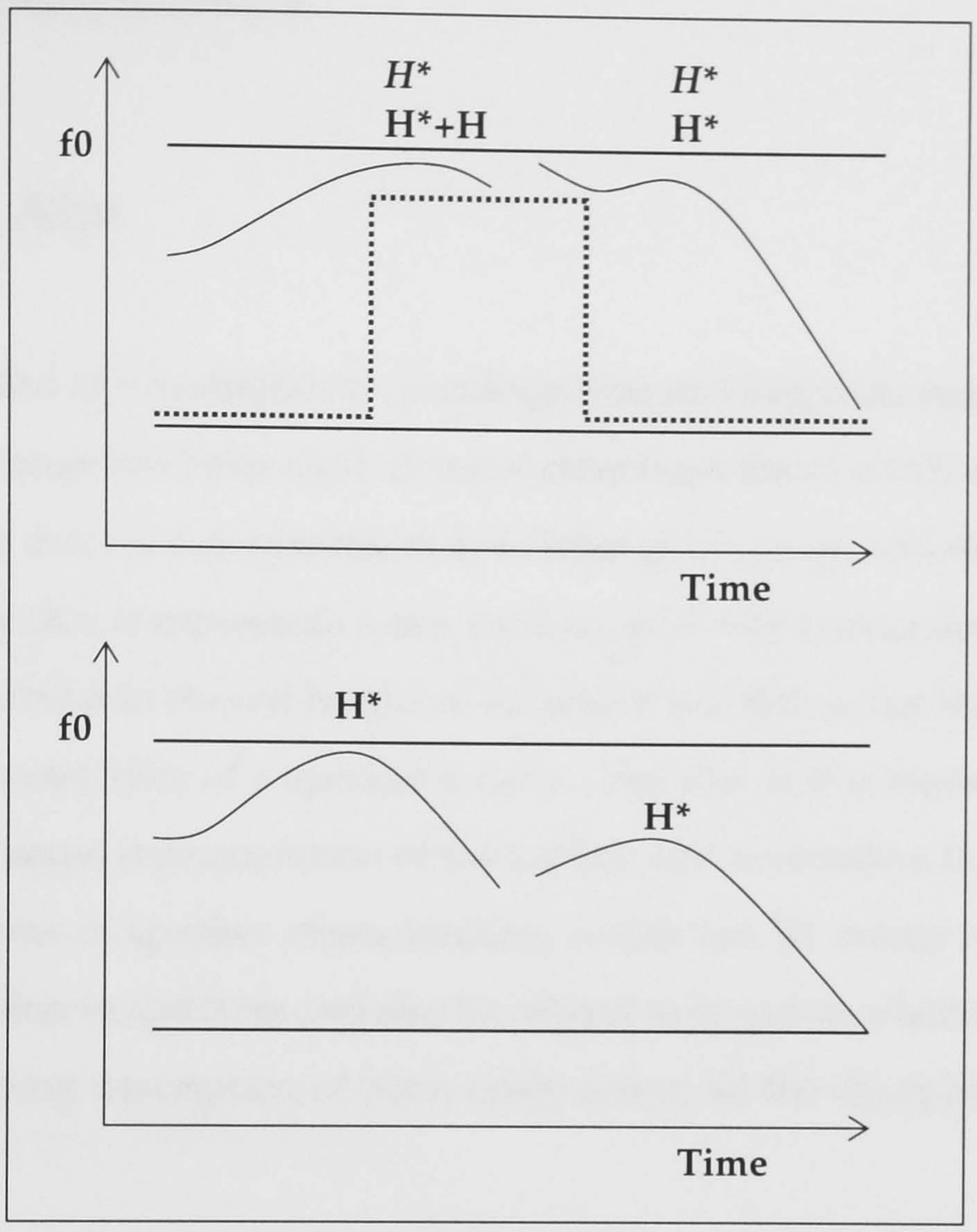



**Figure 1.4.** Schematic drawing of the two contours concerning the H*+H analysis

(1986), which is shown by the italicised accent type, would require a massive leap in the baseline as shown by the dashed line for the top contour. Given that within intonational phonology the H*+H accent type is theoretically implausible it is clear that the two accent types need to be explained in a more theoretically plausible way. The idea that pitch range can be raised and compressed for the purposes of sustaining a high f0 level between two H* accents is not the answer. Given the treatment of pitch range introduced in Liberman & Pierrehumbert (1984) as an explanation, it seems pertinent to point out again how pitch range is defined by Liberman & Pierrehumbert. Pitch range refers to, "a global, or at least phrase-sized, choice of pitch-scaling parameters." The stretch of speech that encorporates the "+H" element of the old H*+H accent is neither global or phrase-sized, therefore the reanalysis, though seemingly convenient,

should be considered unsuitable.

## 1.5 Thesis Aim

There is a long line of extralinguistic, paralinguistic and linguistic research in which a notion of pitch range has been used. It is not clear from the research discussed in this chapter whether there is any agreement as to what pitch range actually refers to other than it is agreed that it represents some topline and some bottomline of a speaker's voice, that a description should be global in nature and that a description should explain some characteristics of a speaker's voice. The aim of this thesis is to see if it is possible to find some representation of the topline and bottomline that can represent a speaker in terms of speaker characteristics, which can be manipulated to explain emotional variation in speakers and also be related to linguistic phenomena, therefore acting as a unifying description of pitch range across all the disciplines discussed in this thesis.

# Chapter 2

# Theoretical and Methodological Issues

## 2.1 Measuring Pitch Range

There are two distinct problems that need to be addressed when establishing a measure of pitch range. Firstly there is the problem of *what* to measure. This problem has been approached throughout the previous chapter, and we will continue to look at it in more detail in this chapter. The second problem concerns the issue of *how* to measure pitch range. There are a number of scales that can be used to measure pitch range, and this issue will also be looked at in more detail in this chapter. Then we will show how we aim to try and solve these problems experimentally.

### 2.1.1 What to measure?

Following Ladd (1996), two independent measures are needed to establish a speaker's range: one which characterizes whether a speaker has a high or low voice, and another

which accounts for whether a speaker's pitch covers a wide or narrow range of frequencies. Various terms have been used to describe these two features: Jassem (1971), for example, uses the terms *pitch* and *compass* respectively. In this study the "height" of a speaker's range will be referred to as the speaker's *level*, while the width of pitch frequencies that the speaker covers will be referred to as the speaker's *span*.

An individual speaker's pitch span and level can, in theory, vary independently of each other. The level of a speaker's pitch can move up or down while the span remains constant, or the span can be expanded or contracted while the level remains constant. This type of variation is exemplified in figure 2.1. In practice these two types of variation do interact: generally speaking, an increase in span is also accompanied by a raising of the level (Ladd & Terken 1995). Figure 2.1 shows that span increases from the bottom up as first shown by Huttar (1968) (cf section 1.3.2). This is one of the 3 theoretically possible ways span could increase, as shown in figure 1.1.

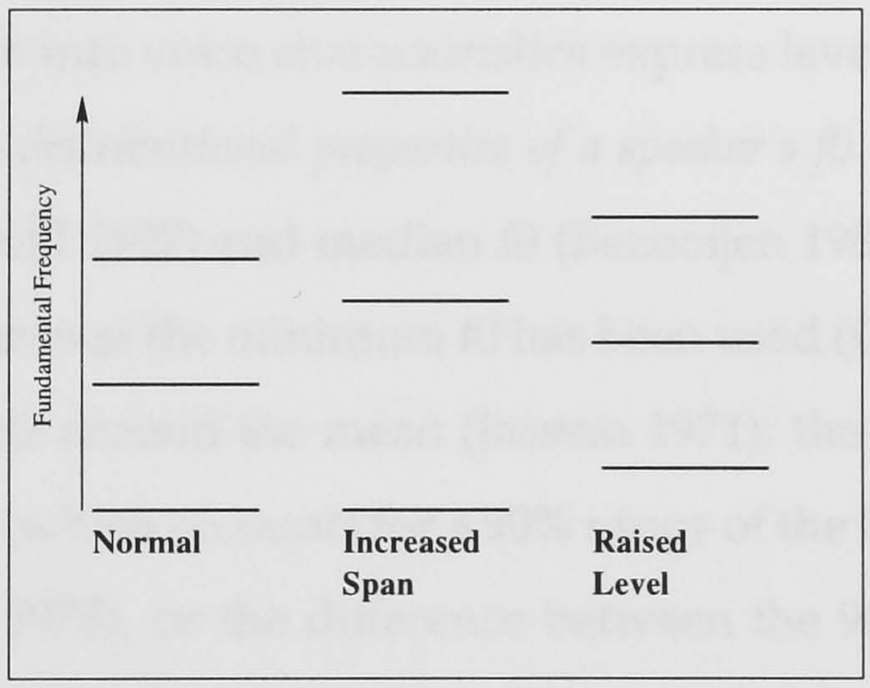



**Figure 2.1.** Possible variations in span and level measures

The independence of level and span are not only necessary to describe within speaker variation. The features level and span can also be used to describe cross-speaker differences. For example, if we allow binary distinctions of level (low/high) and span (narrow/wide), again in theory we should be able to describe four different recognizable voice types; low/narrow, low/wide, high/narrow and high/wide.

A number of different measures of span and level have been proposed. An example of a comparison of cross-speaker characteristics (based on Ladd 1996, p260) illustrates the question of what features of the f0 contour to measure. Compare a low male voice with a range of 60-130 Hz with a female voice with a range of 180-350 Hz. Which ever way one wishes to measure level and span, it is clear that the female voice is higher. But compare one male (speaker A) with a range of 100-160 Hz to another male (speaker B) with a range of 80-180 Hz. Using the lower extremes as a measure of level, speaker A would have the higher level, using the upper extremes speaker B would have the higher level, while using the mean of both speakers' f0 as the measure of level would result in the same level for both speakers. What is clear is that the span of the two speakers differs: 60 Hz for speaker A and 100 Hz for speaker B. We shall draw attention to the most common measures of span and level, as these are the ones that we are going to test experimentally to see which measure of span and level best characterises pitch range.

The majority of research into voice characteristics express level and span measures as related to the *long term distributional properties of a speaker's f0*. For level, measures of both mean f0 (Kraayeveld 1997) and median f0 (Bezooijen 1984) have been used. For span, the maximum f0 minus the minimum f0 has been used (Cosmides 1983), as have four standard deviations around the mean (Jassem 1971), the difference between the 95th and 5th percentile (which accounts for a 90% range of the long term distributional properties of f0, Horii 1975), or the difference between the 90th and 10th percentile (which accounts for an 80% range, Williams & Stevens 1972). An attractive aspect of using long term distributional properties of f0 as a measure of span and level is that it is data that is easy to obtain computationally.

There are, however, problems in using long term distributional properties of f0. For example there are often many spurious measures taken during pitch tracking, including octave errors, especially towards the end of an utterance. These measures may well affect results. Apart from the practical problem of extracting general distributional properties of f0 from a speech signal, there is also a theoretical issue that brings

into question the use of this data to characterise speakers' pitch range. Using a measure of span based on f0 distribution around the mean suggests that f0 is normally or near normally distributed around the mean. Given the data of the speakers recorded for experiments to be reported on in this thesis, the assumption that f0 is distributed normally around the mean is simply incorrect. Nor is it possible to say that there is a clear pattern of skewness around the mean. Patterns of f0 around the mean are very much speaker specific, ranging from normally distributed, postively skewed or negatively skewed. Examples showing the variation in long term distributional properties of speakers can be found in table A.4 through to table A.7 in appendix A.

An alternative to measuring span and level in terms of long term f0 distribution is to see them as fundamentally linked to *tonal targets* found in speech. Tones can be defined as *"abstract elements in terms of which pitch features may be specified: eg abstract highs and lows specifying rises and falls."* (Cutler & Ladd 1983:145). There are two main pieces of evidence suggesting that f0 targets are the phonetic manifestation of underlying static tones (i.e. that f0 contours are structured at a phonological level). First, f0 targets have been shown to be aligned with the segmental string with great consistency across speakers within the same language and dialect group (Arvaniti *et al.* 1998). Speakers have also been shown to have very regular patterns of f0 level at particular points in utterances (Maeda 1976, Liberman & Pierrehumbert 1984).

Bruce & Gårding (1978) suggest that tones are identified with turning points in the f0 contour, such that local maxima correspond to High (H) tones and local minima to Low (L) tones. Pierrehumbert (1980) has suggested that this simple definition is too restrictive: turning points and tones do not necessarily equate in a one-to-one mapping[1]. However, given that there is not a full understanding of the relationship between phonological tones and f0 targets, for the purposes of this thesis we will assume the simple definition proposed by Bruce & Gårding. That is we will assume that turning points in an f0 contour are linked directly to phonological tones and are

[1] A review of both these approaches and issues surrounding them can be found in Ladd (1996:103-105).

therefore linguistic in nature.

Ladd & Terken (1995) conducted a large scale study of pitch range variation, both within- and across-speaker. The findings reported in their paper have been developed further by Shriberg *et al.* (1996). These two papers investigate the relations between global within- and across-speaker differences (both extra and paralinguistic), and the relation of these differences to the more linguistic sources of variation in the scaling of individual pitch targets. The basis of this work rests on the existence of relatively invariant pitch targets in intonation contours, normally local maxima or minima, as previously discussed.

The Ladd & Terken and Shriberg *et al.* corpus consists of seven male and eight female adults, speaking standard Dutch. The speakers were asked to say several sets of sentences designed to elicit specific intonation patterns. In this corpus there are ordinary statements of differing lengths, statements with explicit contrasts, short questions and a news bulletin containing eight short paragraphs. The speakers were recorded in a "normal" situation, then the speakers were instructed to speak as they would if they were having difficulty being heard, (they were told to imagine a poor overseas telephone connection). Full details of these speech materials and further details of the experimental procedures used by Ladd & Terken (1995) and Shriberg *et al.* (1996) can be found in section 3.2.2.

In their study, Ladd & Terken (1995) selected comparable pitch contours from speakers' multiple repetitions of utterances and measured f0 at predetermined points on these contours. They used mean pitch values for their predetermined targets and investigated the patterns of variation in these targets depending on variations such as "raising the voice" and raising due to local emphasis. From their study, Ladd & Terken (1995:388) conclude that:-

- "There is a clear distinction between overall raising and local emphasis. The former raises both peaks and valleys, whereas the latter affects only peaks... In

overall raising of the voice, it is primarily level that is affected. In local emphasis, level is unaffected, but the width of the tonal space is expanded.

- .. it appears that overall raising also slightly raises the speaker's final f0 low.

- For all targets, the effect of raising overall pitch range is extremely constant. For all speakers the correlation between targets in normal range and corresponding targets in raised range is extremely high ( on the order of r = .90)."

Shriberg *et al.* (1996) report that the raising in f0 targets from a neutral mode to a "raised" mode is clearly predictable using a linear function with speaker specific parameters. In slightly more detail, Shriberg *et al.* (1996) establish the relationship between normal and raised targets by examining scatterplots for each speaker in which the mean normal value was plotted against the mean raised value, for all target types and sentence types. The results show that a linear relationship holds between the f0 in the raised mode and the f0 in the normal mode, with the exception of the sentence final low target point. The final lows were excluded from further analysis. Shriberg *et al.* (1996) propose a model with the specific aim of seeking a "raising function" relating the tonal targets in the normal speaking mode to the corresponding targets in the raised mode. This proposal assumes that, to a great extent, speakers have control of their pitch range and deliberately raise it when asked to "speak up". A two-parameter model predicting the raised target (R) from the normal target (N) using a simple linear function was initially used:

$$R = aN + b$$

In this equation, $a$ and $b$ are free parameters, the former accounting for the expansion in f0 span and the latter allowing for any shift in the f0 level (relative to the minimum normal f0). The difference between expansion in span, and shift in level has already been discussed with the example schematically represented in figure 2.1.

Further models were evaluated by Shriberg *et al.* (1996), although none had the same level of accuracy in results. One such model was designed to reduce the number of parameters for the set of speakers by attempting to capture any cross-speaker relationships. Shriberg *et al.* (1996) examined the possibility of there being a universal relationship across speakers that would predict one of the free parameters (*a* or *b*) from the other. They took the results for *a* and *b* from the 2 parameter model and plotted them against each other which showed a roughly linear negative relationship between the 2 variables, i.e. the more level was raised the less span was expanded. The *a/b* points for females lay on a slightly steeper slope than that for males, so it is clear that a cross-speaker model has to take into consideration male/female differences and can not be completely universal. Based on this result each speaker's raising parameters were constrained by the function:

$$b = l * a + m$$

where $l$ and $m$ are now speaker independent parameters, and the values for $l$ and $m$ must be different depending on the sex of the speaker. Shriberg *et al.* (1996) sum up the value of this second model: "Although the one-parameter tied linear model cannot match results for the two-parameter linear model, the tied model is more attractive from a theoretical point of view, since it directly reflects similarities as well as differences across speakers. In addition, the tied model may be preferable from an applied perspective, since it reduces the overall number of parameters to be estimated."

Returning to the link between tonal targets and pitch range, if one is a supporter of linguistically motivated dimensions of variation in pitch range, then it is still not clear which linguistic targets best characterise a speaker's span and level. There are a small selection of justified potential candidates to be considered. In principle, sentence final low is a suitable measure of level as it is considered the most stable of targets (Maeda 1976). On the other hand, this is a low target that is in *isolation* compared to all the other valleys found in f0 contours. By isolation, we mean that it is different

to the other lows and is therefore uncharacteristic of "low" in a more general sense. Research has shown that sentence final low may not be as unaffected a target to pitch range modifications as initially considered (Hirschberg & Pierrehumbert 1986, Ladd & Terken 1995). Having suggested that sentence final low is not characteristic of low in a general sense, it is necessary to consider a measure that is more characteristic of low found in a speaker's f0 contour. We propose that level could best be characterised as an average of a speaker's post-accent valleys, as these more readily appear in any speaker's f0 contour.

Span can be defined as the difference between a certain topline and a certain bottomline. I've already identified two possible bottomlines: an average of a speaker's sentence final lows and an average of a speaker's post accent valleys. These will also double up as the possible measuring points for level. There are also two potential and distinct toplines: an average of a speaker's sentence-initial high or an average of all of a speaker's non-initial accent peaks. In a study of pitch variation in read speech (from the Boston Radio New Corpus (Ostendorf *et al.* 1995)) Clark (1999) found that the first tone group[2] in a phrase has a greater pitch range and a higher mean than any other tone grouping. For the one speaker analysed in the Clark (1999) study, the f0 mean of the non-phrase-initial tone groups is around 165-170 Hz, whereas the mean of the initial tone groups is around 200 Hz. This suggests that the first tone group has some special status. The phrase-final groups are slightly lower than the other categories, but not to the same extent to which the phrase initial tone groups are higher. Clark also shows that the medial tone groups all appear to be very similar in their characteristics. This impressionistic view is supported by statistical analysis showing that initial and final tone groups differ from each other, and from medially positioned tone groups, but that all medially positioned tone groups are effectively the same.

Given the support for invariant pitch targets in intonation contours (Maeda 1976, Liberman & Pierrehumbert 1984, Arvaniti *et al.* 1998), and given the predictability

[2]For the purposes of the Clark study, a tone group was defined as a group having a ToBI break index of at least 3. For an introduction to the ToBI transcription see Silverman *et al.* (1992).

in their variation when modified by the paralinguistic effect of speaking up (Ladd & Terken 1995, Shriberg *et al.* 1996), it seems reasonable to see how these linguistic targets could be used to measure pitch range for speakers' voice characteristics. We believe that pitch targets in speech will define an appropriate phonetic description of the phenomenon of pitch range. The aim of this thesis is to show that a model of pitch range can be successfully based on these linguistic tonal targets, and can be used to better characterise speaker characteristics. Therefore we have strong support for unifying a model of pitch range that suits the needs of all the research strands discussed in chapter 1. This will also give support to the belief that, "paralinguistic cues should be regarded as *modifications of the way in which phonological categories are realised*." (Ladd 1996:35).

### 2.1.2 How to measure?

In speech research to this point, pitch has been expressed in terms of a number of different units. Frequency is generally expressed in terms of the unit Hertz (Hz) which is a linear scale and has been used by Cooper & Sorensen (1981). However, while Hz is a long established unit of measure, it may not entirely be suited to pitch range research. We are obviously looking for the measure of pitch range that characterises a speaker as well as possible. A linear scale such as Hz certainly can capture the differences in level between men and women successfully. A male speaker with a range between 100 and 200 Hz clearly has a lower level than a female speaker with a range between 200 and 400 Hz, whichever measure of level is taken. But to say that the male speaker has exactly half the span of the female speaker (100 Hz as compared to 200 Hz) could well be misleading: the nature of the auditory system will not rank the male speaker's span as being exactly half that of the female span. The linear Hz scale may therefore not be characterising the pitch range successfully.

Investigations in hearing research have used other scales of measurement, like the well established musical scale of semitones (e.g. 't Hart *et al.* 1990). The musical scale

is logarithmic, in which equal distances between two tones represent equal frequency proportions. Returning to the example of the male and female speaker above, the 100 Hz span between 100 and 200 Hz represents the same distance on a musical scale as the 200 Hz span between 200 and 400 Hz for the female speaker. On a musical scale, the level of the female speaker is again clearly higher, but the span between the two speakers is the same. In a review of scales used in speech research Hermes & van Gestel (1991) cite Graddol (1986) who states that, "whenever intervals in pitch must be compared at different frequencies, a log scale is to be preferred." Given that the musical semitone scale is logarithmic, it is suitable for measuring span, as opposed to level. Semitones represent distances between two tones, so for span this would be the difference between the topline and the bottomline. Semitones (st) can be calculated from Hz by the following formula in which st is the number of semitones between frequencies f1 and f2 (where f1 is the bottomline and f2 is the topline):

- $st = \frac{12.ln\frac{f2}{f1}}{ln2}$

In psychoacoustics, a number of scales have been derived from the frequency selectivity of the auditory system, including the Mel scale (Stevens *et al.* 1937), the Bark scale (Zwicker 1961) and the equivalent-rectangular-bandwidth-rate (ERB-rate) scale (Patterson 1976). Recent work has used the ERB-rate scale for describing the size of pitch movements (Hermes & van Gestel 1991, Hermes & Rump 1994, Shriberg *et al.* 1996), so we shall go into a brief look at the derivation of the ERB-rate scale, and how it has been applied.

The basilar membrane is a membrane inside the cochlea which vibrates in response to sound and whose vibrations lead to activity in the auditory pathways. It is this part of the peripharal auditory system that is described as being equivalent to a bank of bandpass filters (Helmholtz 1954). A bandpass filter has two cutoff frequencies, passing components between these two frequencies (known as the bandwidth), and removing components outside this range. So each bandpass filter in the ear has a

different centre frequency so that the whole range of audible frequencies is covered.

One measure of bandwidth is the equivalent rectangular bandwidth (ERB). Moore (1997) provides a definition of the ERB measure of bandwidth. "The ERB of a given filter is equal to the bandwidth of a perfect rectangular filter which has a transmission in its passband equal to the maximum transmission of the specified filter and transmits the same power of white noise as the specified input." While the derivation of the ERB is designed to describe the frequency selectivity of the auditory system, based on the perception of a signal through noise, Moore (1997) goes on to say that, "sometimes it is useful to plot psychoacoustic data on a frequency scale related to the ERB. Essentially, the ERB is used as the unit of frequency."

Dik Hermes and colleagues (Hermes & van Gestel 1991, Hermes & Rump 1994) have argued that the most appropriate scale for measuring f0 is the ERB-rate scale. They studied the role of the size of pitch excursion to the perceived prominence of a syllable in a spoken context. They tried to find out on which scale a pitch excursion in a low (male) voice must be equal to a pitch excursion spoken in a high (female) voice in order to lend the same prominence to a syllable. Their experiments showed that the ERB-rate scale is the appropriate scale for this. The consequence of this is that, if two intervals, one spoken in a low register, and the other in a high register, are equal on a linear frequency scale, i.e. a Hz scale, the syllable accented by the pitch movement in the high register will be perceived as less prominent than the syllable in the low register. For another example, if two intervals, one spoken in a low register, and the other in a high register, are equal on a log frequency, i.e. a musical scale, the syllable accented by the pitch movement in the high register will be perceived as more prominant than the syllable in the low register.

Continuing the example of the two speakers used above, if the male speaker raises his pitch from his minimum 100 Hz to his maximum 200 Hz, a question that researchers need to answer is by how much would the female have to raise her pitch from her minimum 200 Hz to match the same pitch excursion of her male counterpart. Raising

on a linear scale would mean a matched increase of 100 Hz, to 300 Hz. Raising on a logarithmic scale would mean an increase of 200 Hz to 400 Hz. On an ERB-rate scale, the increase would be of approximately 139 Hz, from 200 Hz to 339 Hz. The fomulae used in this thesis for conversion between the linear and the psychoacoustic scale are taken from Hermes & van Gestel (1991) where f is frequency in Hz, and E is the ERB-rate in ERB:

- $E = 16.7 \log_{10}(1 + f/165.4)$
- $f = 165.4(10^{0.06E} - 1)$

The results of Hermes & van Gestel (1991) for speech are in contrast to what is found in music, where intervals between notes are equal if they are equal on a logarithmic (e.g. a semitone) scale. Apparently, the perception of pitch is essentially different in speech and in music.

There are various other essential differences in the perception of the melody in speech and in music. For instance, in music there is a limited number of correct notes within an octave. The pitch of a note can be too low or too high, with respect to the target. In speech there is only continuous change from lower to higher and the pitch of a syllable, even if perceptually clear, cannot be too low or too high. If excursion sizes are too large or too small, the perceived prominence of a syllable may be too large or too small, but there is no such thing as a wrong note.

The results of Hermes & van Gestel (1991) are not without controversy. Traunmüller & Eriksson (1995) are skeptical about Hermes & van Gestel's choice of considering only the first partial (f0). Traunmüller & Eriksson (1995) cite Ritsma (1967) who showed the third, fourth, and fifth harmonic to dominate pitch perception. Traunmüller & Eriksson (1995) note that the problem of deciding which partial to consider does not arise if pitch is scaled logarithmically. If expressed in semitones, the excursions of all the partials are the same whereas if they are expressed using ERB, Hz (and the other

psychoacoustic scales Bark and Mel which we have not discussed), the excursions of all the partials are different.

Clearly, the issue of scale is still not resolved and for the purposes of this thesis, investigation will be made into all three types of scale to find which one best captures differences in pitch range.

### 2.1.3 Examples of variation in measuring pitch range parameters

Having described the variations found in what to measure and how to measure pitch range, it would be useful to show an example of the variations in pitch range measurements that can be found. Figure 2.2 and figure 2.3 show the f0 contour for two male speakers saying the sentence, *"What am I going to write?"*.

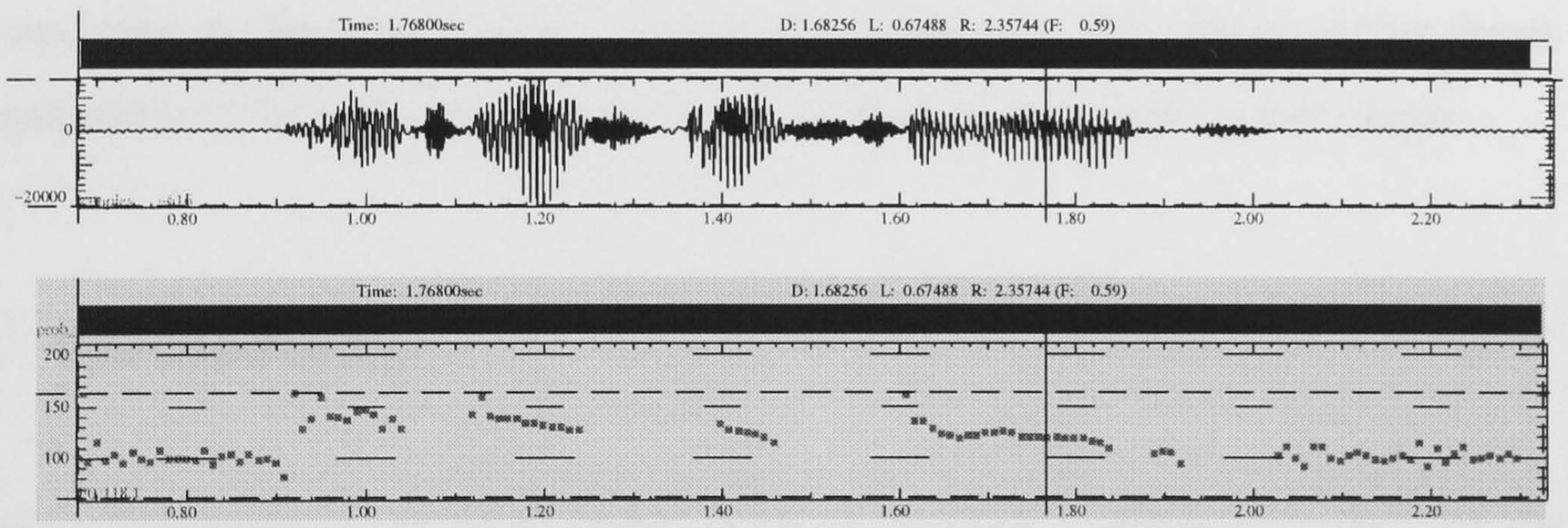



**Figure 2.2.** Speaker A: What am I going to write?

Just from inspecting the f0 track for both speakers it is clear that speaker A has a narrower span than speaker B. It would appear that speaker A has a lower level as well, but it is not entirely clear by how much because the pitch track towards the end of the utterance is a bit messy for both speakers. These impressionistic results are verified by a whole variety of measures as shown in table 2.1. Speaker A always has a lower level and a narrower span than speaker B, though by what extent depends on the version of the pitch range measure choosen. For reference, the features mentioned

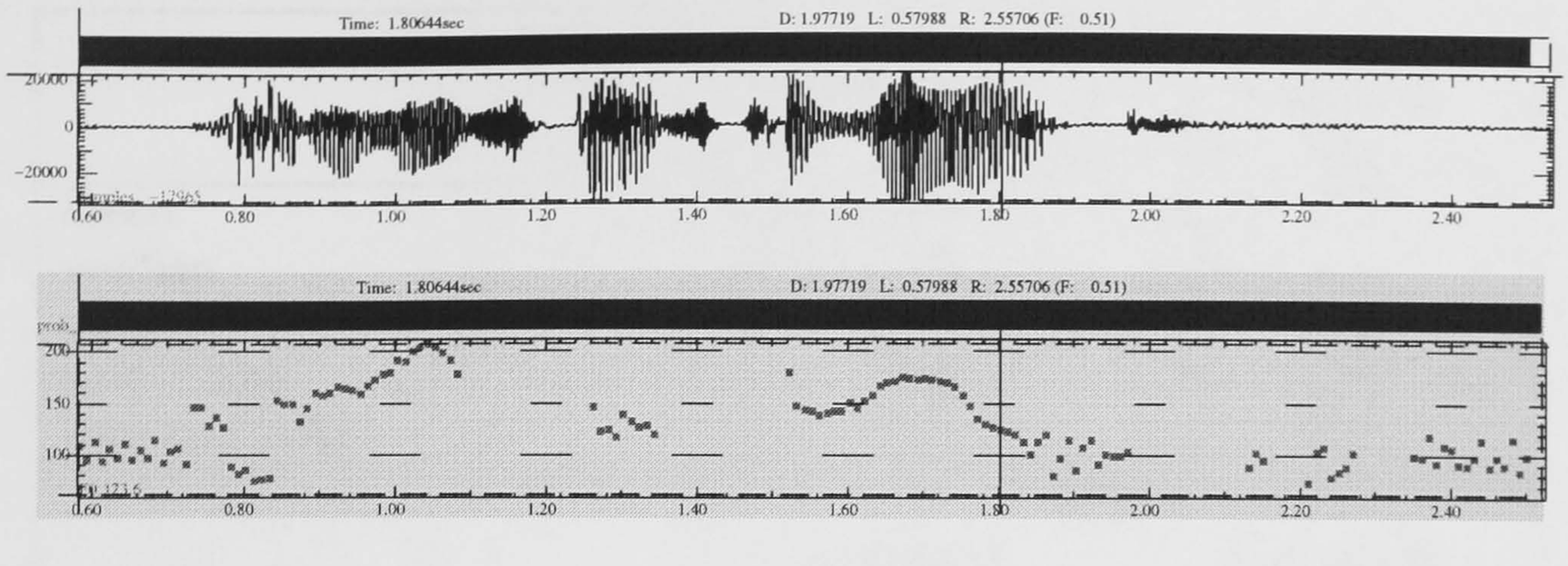



**Figure 2.3.** Speaker B: What am I going to write?

for span and level in table 2.1, and which have all been mentioned previously are, in order, mean f0, median f0, minimum f0, sentence final low, post accent valley, standard deviation, 4 standard deviations around the mean, maximum - minimum, the 95th - 5th percentile, the 90th - 10th percentile, sentence-initial high - sentence-final low and non-initial-accent high - post-accent valley. The remainder of this thesis will investigate which of the measures in table 2.1 best characterises pitch range.

## 2.2 Methodology

What we require is a valid and reliable method of judging which measures of level and span are the most effective at characterising a speaker. Reliability is usually thought of as the degree to which measuring instruments of the same types give the same results. A large literature has developed concerning the causes, effects and remedies for deficiencies in reliability (eg Scherer & Ekman 1982). In general, present-day social science shows a high degree of sensitivity to problems of reliability. Validity, like reliability, involves agreement between measures. Validity, however, involves agreement between maximally different, or independent, measurement procedures; whereas reliability involves agreement between maximally similar measures. The two different methods producing very divergent results cannot both be measures of the same thing

| | LEVEL | | | | | |
|---|---|---|---|---|---|---|
| | Speaker A | | | Speaker B | | |
| | Hz | ERB | s.t. | Hz | ERB | s.t. |
| mean | 120.4 | 3.97 | | 131.7 | 4.25 | |
| median | 119.2 | 3.94 | | 135.9 | 4.35 | |
| min | 81.7 | 2.91 | | 55.4 | 2.10 | |
| sent. f. low | 101.2 | 3.46 | | 93.8 | 3.26 | |
| post acc. valley | 114.4 | 3.81 | | 116.6 | 3.87 | |
| | SPAN | | | | | |
| | Speaker A | | | Speaker B | | |
| | Hz | ERB | s.t. | Hz | ERB | s.t. |
| s.d. | 20.8 | 0.86 | 1.19 | 37.9 | 1.49 | 1.34 |
| ± 2sds mean | 81.0 | 2.89 | 2.01 | 141.8 | 4.49 | 3.33 |
| max - min | 80.7 | 2.88 | 1.99 | 150.1 | 4.68 | 3.71 |
| 95th - 5th percent | 65.3 | 2.41 | 1.68 | 122.3 | 4.02 | 2.67 |
| 90th - 10th percent | 61.3 | 2.29 | 1.63 | 101.6 | 3.47 | 2.32 |
| sent. i. high - sent. f. low | 44.0 | 1.71 | 1.43 | 111.4 | 3.73 | 2.19 |
| non i. acc. high - post acc. vall | 20.1 | 0.83 | 1.18 | 56.2 | 2.12 | 1.48 |

**Table 2.1.** Variation in level and span measures

although both might be valid measures of different things.

Scherer & Ekman (1982:169) says that,*"Almost all of the objective parameters of an acoustic speech wave form can be 'heard' by judges and can consequently be assessed with the help of category systems and rating scales."* A clear and obvious method to test which measure best characterises range phenomena is to run a perception experiment to try and capture what listeners actually hear. At first sight the simplest way to measure which span and level measure is the most effective is to play the speech of many speakers to a group of listeners, asking them to assess the pitch range of each speaker. After doing this, a comparison could be made between the various acoustic measures of pitch range and the perceptual measure of pitch range. This certainly is the most direct method for testing the acoustic measures of pitch range, if one assumes that there should in theory be a strong correlation between the acoustic and perceptual measures of pitch range.

There are experimental flaws to be found in this method though, and we would suggest that this most direct route to pitch range assessment is not valid or reliable. Scherer & Ekman (1982:169) goes on to say, *"However, owing to a number of factors, these auditorily assessed variables do not necessarily correlate very highly with objectively measured variables of the same acoustic parameters. Among these factors are the characteristics of the human hearing system (which does not function exactly like a filter bank analyzer or a digital computer); the fact that expectations and auditory habits, often based on certain aspects of a language and/or cultural norms, affect auditory impression; and the difficulty of translating an auditory impression into a quantitative judgement on a scale."*

The main issue of this thesis is to examine the possibility of unifying accounts of pitch range across linguistics and voice research. As is clear from section 2.1.1, there are many variations as to how pitch range should be characterised acoustically. In a pilot experiment, 3 experienced phoneticians were asked to rate 40 speakers on a host of voice quality criteria as well as to categorise the pitch span of each speaker as neutral, narrow or wide. Of the 40 speakers, results showed that speakers could only agree on the category of span for 12 speakers. Therefore it is considered that asking naive listeners to make pitch range judgements most certainly would be unreliable.

The methodology chosen in this project is a more indirect method to assess the measures of pitch range, and uses as its basis the long tradition of work found on pitch range and paralinguistic communication. The methodology to be outlined here, and given in more detail in the following experimental chapters of this thesis, may be more indirect, but more valid than the more obvious direct method outlined above. The reliability of the methodology is not taken for granted though, and is tested as well.

The proposal is to record a number of test passages spoken by as large a number of speakers as possible. A panel of listener judges would be asked to rate each speaker on a number of phonetic and pragmatic criteria. Given the long line of research connecting range to speaker characterisitics, the strength of correlations between subjects'

judgements and the variations in pitch range could then be examined. The method of measuring span and level, and the scale of measurement could then be assessed. The method with strongest correlations most effectively characterises cross-speaker differences.

It has been acknowledged that there are a number of factors (e.g. voice quality, loudness, pitch span, pitch level) which contribute to a listener's overall perception of a speaker's "character". While it is accepted that no one discrete characteristic can be explained by one of these voice factors alone, there is a wealth of research that relates the independent contribution of pitch range to a class of character types (Scherer 1979, 1981, Scherer & Ekman 1982, Ladd *et al.* 1985). The general pattern of results from studies shows that the wider the pitch span the more positively speakers are characterised. Uldall (1964) describes these positive attributes as being on a scale of pleasantness, while Brown *et al.* (1973) patterns these positive attributes on a scale of competence. Results from Ladd *et al.* (1985) show that pitch level is strongly correlated with arousal. The independence of pitch range in the perception of speaker characteristics has been supported experimentally using masking techniques (e.g. low-pass filtering cf. section 4.4) to block other voice features from affecting results (Scherer & Ekman 1982). Given that there are clear findings showing the independence of pitch range in the perception of speaker characteristics, it is perfectly reasonable to suggest that judgements of speaker characteristics can shed light on measures of pitch range.

Having discussed the issues of pitch range, and the various details on how to measure range, we have suggested a methodology for assessing pitch range using speaker characteristics. Therefore it is necessary to have a look at research done on collecting such data, and justifying the validity of our research design and research techniques.

### 2.2.1 Which type of speech to study?

There is some debate as to what the best type of speech material to use when investigating issues like the conveyance of emotion in speech and assessing speaker characteristics. It would seem reasonable to suggest that the best material would be that produced spontaneously in a real-life setting. It is clear however that relatively few empirical studies on vocal expressions of emotion have made use of natural speech, a notable exception being Williams & Stevens (1972). A problem with such analysis is the verbal content of the speech samples is not controlled for. Another problem in using spontaneous speech is that it is hard to know what emotion the speaker was experiencing when producing it. The majority of research uses simulated emotion produced by actors, which would seem to bring along with it just as many problems. Greasley *et al.* (1996) report on a series of four experiments in which subjects listened to ninety-one episodes of emotional speech. Subjects were required to judge the emotions expressed by:

- using a word of their own choice.
- selection from a list of 22 emotion types.
- selection from a list of 5 basic emotions.

Analysis of the results showed that naturally occurring emotional speech presents a much more complex picture of emotion perception then that found in studies using actor portrayals of emotion.

As discussed in chapter 1, Leinonen *et al.* (1997) have studied the emotional variation found in the one-word Finnish utterance [saara]. The motivation behind the choice of a one word utterance is based on the idea that the expression of emotion will be exaggerated as well as the hope of reduction of inter- and intra-subject variation. This would facilitate the identification of meaningful signal dimensions. Certainly exaggerated speech can be used to make significant research contributions but it must

also be acknowledged that emotions are still clearly expressed in natural, "everyday" speech as used by Huttar (1968) and by Scherer *et al.* (1984), for the same reasons. Generally speaking, there is not often confusion in picking up whether someone is confident or moody, and in order to fine-tune the necessary details in, for example, speech synthesis, it is important not to place too much importance on the results of exaggerated speech.

Uldall (1960, 1964) looked at the emotional meaning of selected contours. She gave subjects a number of sentences on each of which 16 intonation contours had been imposed synthetically. The sentences were intended to be as colourless as possible so as to allow the intonation to add as much as possible to their meaning. The intonation contours varied along three parameters:

- range: wide/narrow
- pitch reached at end of contour: high/mid/low
- shape of contour: one direction/with a change of direction

Scherer (1981:204) says, "If the speech samples used in simulation studies are not natural enough, the samples consisting of spontaneous speech are not emotional enough." Scherer (1981:205) also says, "*Despite the large number of methodological flaws in many relevant studies, the patterns of results is surprisingly consistent, testifying to the stability and strength of emotion effects on voice and speech.*"

Given the selection of speech options available - natural, acted or synthesised - this thesis reports on speech data collected from recordings of speakers reading controlled texts in a recording studio, in as natural and as comfortable way as possible. No emotions were asked for as it is believed that there are enough ratable differences between speakers without the need for any emotional performance. Although more realistic speech might be preferable in theory, the need for high quality recordings for acoustical analysis was more important.

### 2.2.2 How to measure speaker characteristics?

The two generally accepted methods available for the measurement of speaker characteristics are the Category Rating (CR) and the Magnitude Estimation (ME) methods.

The CR procedure is the most commonly used method for scaling or measuring strength and direction of subjective states (Lodge 1981). CR methods include methods such as Likert scales; the prototype is the familiar row of boxes with a label at each end to define the continuum, such as "weak - strong" or "not at all confident - very confident". Their great advantage is their simplicity; their great disadvantage is the fact that they do not produce estimates of the perceptual magnitudes of the stimuli used, but rather produce estimates of the relative discriminability of the stimuli, which yield, for most purposes, much less useful information. The measurement of these discriminabilities is at an ordinal level as opposed to interval level data which if possible to obtain would be more desirable.

It is possible to distinguish at least four kinds of scale, which are called nominal, ordinal, interval and ratio (see table 2.2, from Stevens 1974). Depending on what type of scale we have constructed, some statistics are appropriate, others not . The group of mathematical transformations permitted on each scale determines which statistical measures are applicable. In general, the more unrestricted the permissible transformations, the more restricted the statistics. It should be noted that ordinal scales have the property of order or sequencing of scale values, but do not have a meaningful (or unique) origin. Although the distance between values is not known with an ordinal scale, ordinal scales are often treated statistically as though they in fact possess interval level properties; e.g. the unit sizes between values are treated as though they are equal.

ME procedures avoid the use of preset rating categories. Instead instructions are given which emphasize that responses should be proportional to the intensity of experienced subjective states. Stevens (1957) proposed that ME procedures constitute

| Scale | Basic Emperical Operations | Mathematical Group Structure | Typical Examples |
|---|---|---|---|
| Nominal | Determination of Equality | Permutation Group<br>$x' = f(x)$ where *f(x)* means any one-to-one substitution | "Numbering" of football players<br>Assignment of type or model numbers to classes |
| Ordinal | Determination of greater or less | Isotonic Group<br><br>$x' = f(x)$ where *f(x)* means any increasing monotonic function | Hardness of minerals<br><br>Street numbers, grades of leather, lumber, wool, etc. Intelligence test raw scores |
| Interval | Determination of the equality of intervals or differences | linear or affine group<br><br>$x' = ax + b, a \rangle 0$ | Temperature (Fahrenheit or Celsius)<br><br>Position, Time (calendar), energy (potential), intelligence test "standard scores" |
| Ratio | Determination of the equality of ratios | Similarity group<br><br>$x' = cx, c \rangle 0$ | Numerosity, length, density<br><br>work, time intervals, etc. Temperature (Kelvin), Brightness (brils) |

**Table 2.2.** A Classification of Scales of Measurement

"direct measurement" of sensations, or subjective experience. Stevens proposed the Power Law (a power function equation) to represent the relation between the strength of the sensation and the strength of the physical stimuli. According to the Power Law, equal stimulus ratios produce equal subjective ratios. Expressed in the form of an equation, the Power Law is:

$$Y = kX^b$$

where $Y$ is the sensation or subjective magnitude of the experience, $k$ is a constant of proportionality, $X$ is the actual physical magnitude of the stimulus and $b$ is the value of the exponent which characterizes the relationship between objective and subjective stimulus magnitude. The term "direct measurement" proceeds from the assumption in the Power Law that there is a constant, linear relationship with a zero intercept, between the reported magnitude estimation of the subjective state ($Y$) and the actual subjective magnitude of that subjective state. Thus it is claimed that the magnitude estimation produces ratio scale measurement.

When making magnitude estimations judges attempt to match the magnitude of a number to the magnitude of the sensation produced by a stimulus magnitude. Stevens (1957), who initially developed the procedures, left the following as rules:

- Use a standard whose level does not impress the observer as being extremely soft or extrememly loud (i.e. use a standard in the middle of the stimulus range).
- Present variable stimuli that are both above and below the standard.
- assign a number to the standard only, and leave the observer completely free to decide what he will call the variable. In particular do not tell the observer that the faintest variable is to be called "1" or that the loudest is to be called some other number. (If the experimenter assigns numbers to more than one stimulus, he introduces constraints of the sort that forces the observer to make categorical rather than magnitude judgements).

- Use only one standard in any experiment, but use various standards in later replications, for it is risky to decide the form of a magnitude function on the basis of data obtained with only one standard.

- Randomize the order of presentation. With inexperienced observers, it is well, however, to start with stimuli that are not extreme and are therefore easier to judge.

- Make the experimental session short, about ten minutes.

- Let the observer present the stimuli to himself. The observer can then work at his own pace and so is more apt to be attending properly when the stimulus comes on.

Increasing attention must be drawn to issues involving the quantification of subjective states. Of recent years the trend has been towards magnitude estimation procedures over the more traditional category rating methods in dealing with judgements of strength or intensity of perception. The comparative benefits, disadvantages and methodological issues associated with the use of magnitude estimation and category rating methods of scaling have been discussed in several recent articles (Meek *et al.* 1992, Sennot-Miller *et al.* 1988).

It has been suggested that ME is superior to category rating methods for the scaling variables which involve judgements of the intensity or strength of perceived stimuli (Lodge 1981, Stevens 1957, Meek *et al.* 1992). However there are psychometric controversies associated with the use of ME which are often neglected. The assertion that ME yields "direct measurement" of subjective states has been criticised for several reasons.

First, the assumption of a linear relationship between subjective state and ME response has generated much debate. Birnbaum & Veit (1974) proposed that the relationship between a subjectively experienced magnitude of sensation, and the response (or rating) on a rating scale may be distorted by the ME procedure, but not by the CR

procedure. This implies that the ME responses do not map onto the sensations in the way that Stevens assumed. Both CR and ME methods can be assumed to produce at least ordinal level scales of subjective states.

A second and related criticism of the "direct" measurement argument is the finding that people use the same mental comparison process to make judgements of sensation strengths, regardless of whether CR or ME methods are used (Birnbaum & Veit 1974). Because the two methods do not agree, one of the two methods must be yielding a response pattern that distorts the sensations. The problem is that without further criteria there is no way of knowing which method yields a distorted measure of sensation, ME or CR.

Lodge (1981) and Stevens (1957) have been cited as major authorites for the argument that ME is superior to CR methods. Lodge (1981) argued that a serious limitation of CR methods is the loss of information due to limited resolution associated with use of preset categories. However, whether or not a particular CR method lacks sensitivity in measuring subjective states depends upon at least two factors. First, the extent to which research participants are able to distinguish or discriminate different levels of the relevant stimulus may vary considerably, both between subjects and as a function of the type of stimulus. "Lack of sensitivity" may just as readily be a function of the rater or the stimulus as the rating scale.

Second, motivational factors may influence the ability or desire of research participants to distinguish among levels of a stimulus or to make ratings of subjective states. Unwillingness to expend cognitive effort may make collecting ratings of subjective states difficult, and may influence how research participants use a rating scale. Under such circumstances, the complexity of ME instructions may sometimes be a disadvantage. It is widely accepted that CR procedures are easy for participants to understand.

A further criticism is that CR methods may fail to include categories that are representative of the full range of subjective values of various stimuli. This disadvantage is a limitation only if the investigator fails to give the participant practice trials which are

representative of the type of stimuli to be evaluated. It has been shown that participants will adjust their use of a rating scale (either category or magnitude estimation) to accommodate the range of stimuli encountered (Mellers 1983). Although this implies that measurements of subjective states by CR are not absolute, such a fact is not a disadvantage of CR compared to ME. Both scaling methods are subject to such effects. For this thesis the CR method was selected. Despite some of the favourable aspects of the ME method, there is a long tradition of using the CR method in speaker characteristics and emotion in speech research. Essentially CR methods are easier to construct, easier to run, easier to understand, and are efficient in the acquisition of a large amount of data, especially in the complex arena of speech studies. All further discussion of methodological issues relate to specific experiments and will therefore be detailed in the forthcoming experimental chapters.

## 2.3 Conclusions

So far it has been shown that there are a number of unresolved issues relating to the characterisation of pitch range. It is unclear whether pitch range should be related to different things depending on the nature of the research, which acoustic measure best characterises within- and across-speaker differences in pitch range and which units of measurement best characterise these speaker differences in pitch range.

# Chapter 3

# Experiment 1

## 3.1 Introduction

This chapter reports on a first experiment relating a linguistic model of pitch range to speaker characteristics using Dutch speakers. The Shriberg *et al.* (1996) study has developed a model for within-speaker variation in pitch range, and it is necessary to see if it adequately captures variation in pitch range across-speakers. The ideal model of pitch range should be able to account for both sources of variation.

An initial research proposal was made in Monaghan & Ladd (1990:8). *"We propose to record a text passage as spoken by 100 speakers. A panel of subjects would be asked to rate each speaker on both phonetic and pragmatic (emotional, attidudinal, etc.) criteria, and the correlations between subjects' judgements and the variations in pitch range would then be examined. This data could then be used to refine the existing model* [described in the paper which later developed into the Shriberg *et al.* (1996) model] *or to evaluate alternative models according to the correspondence of their parameters with subjects' judgements..."*. This experiment can be considered a small scale version of the Monaghan & Ladd (1990) proposal, which until now still remains a proposal.

The aim of the current research is to take the data from the pitch range model and use it along with the speech recordings from the Ladd & Terken (1995) and Shriberg *et al.* (1996) studies to carry out, at least to some degree, the proposal to examine correlations between subjects' judgements of phonetic and pragmatic criteria[1] and variations in pitch range. The aim of examining these correlations is to show that the model sufficiently captures patterns found in previous research (Uldall 1960, Brown *et al.* 1973, Pakosz 1982). Apart from the connections with pitch range, another of the patterns that unites the work of these three papers is in showing the similarities between emotions, similarities that may be conceived as proximities in a multi-dimensional space. An example of a three dimensional model (Pakosz 1982) is:

- Evaluation - positive vs negative (e.g. pleasantness vs unpleasantness)
- Activation - strong vs weak (e.g. horror vs complacency)
- Control - active vs passive (e.g. contempt vs fear)

In all these papers, pitch range has most strongly correlated with the *evaluation* dimension. Finding similar patterns to previous research will establish the methodology that we have chosen as valid. Once this is established it is then necessary to find out whether a model based on Shriberg *et al.* (1996) is in fact better than any other suggestions for measures of pitch range.

The rest of this chapter will fall into two parts; a description of the work that went into the acoustic study of speakers' pitch range and a description of the perception study which was used to collect the judgements of listeners for certain speaker characteristics.

---

[1] These criteria are outlined in section 3.4.4.

## 3.2 Stimulus Design and Analysis

The speech recordings and the pitch data extracted from these recordings, which we used in this initial study based on Dutch speech, were made for the Ladd & Terken (1995) and Shriberg *et al.* (1996) projects previously discussed (page 38). We shall report on all the details necessary and relevant to the materials that we used. The following details relating to the speech recordings are an expanded version of that which can be found in the methodology of the Ladd & Terken (1995) paper. Additional information has come from Ladd (personal communication).

### 3.2.1 Speakers

The speakers were 16 native speakers of Standard Dutch, 8 males and 8 females, all students or employees at the Institute for Perception Research (IPO). None of these speakers were closely involved in research on prosody. None of the speakers knew in any detail what the aims of the initial Ladd & Terken (1995) and Shriberg *et al.* (1996) studies were and they certainly could not have had any idea as to the nature of the current study. From this pool of 16 speakers, only material from 11 speakers could be used for our study due to problems in transport of the data from the speech laboratory at IPO to its counterpart in Edinburgh. In the current study the speech of 5 males and 6 females was used. The speakers recorded are considered to cover a wide range of voices, from a deep male voice (speaker JR) to a high female voice (speaker IS). In the current study there is an assumption that there is enough variation in the voices to be able to make "ratable" differences in speaker characteristics, rather than relying on recordings of acted or simulated characteristics. There is also a clear reason for choosing Dutch speakers above and beyond the fact that it was an easy database of speech to obtain. British listener judges could be encouraged to really focus in on the task in hand i.e. rating the voice without the distraction of the semantics of the utterances. Also it was thought that regional accent would be a biasing factor if

English speech was to be used.

### 3.2.2 Speech Materials

The basic approach used by Shriberg *et al.* (1996) was to measure f0 at specific pre-selected points in multiple repetitions of utterances with comparable contours. By doing this they hoped to establish stable mean pitch values for certain putative target levels (e.g. first accent peak, utterance-final low in statements etc), creating a kind of "map" of the relative pitch of these targets which could then be compared between speakers or between different pitch range settings of the same speaker.

#### Sentence Design

Shriberg *et al.* (1996) designed several sets of sentences intended to elicit specific intonation patterns which they expected would have consistent and identifiable peaks and valleys at well-defined points. These included ordinary statements of varying lengths, short questions and statements with explicit contrasts (of the sort "Not X but Y"). Speakers were also asked to read a news bulletin containing 18 short paragraphs. The recording session also included a short section of spontaneous speech (a description of the speaker's route to work).

The speech materials used for the current experiment were taken from a selection of utterances that fall into three main groups. Each group was intended to contain 32 utterances, either 2 repetitions of 16 sentences or 4 repetitions of 8 sentences. As it happened, some of the test sentences were realised with inconsistent intonation patterns and were excluded from the analysis of the Shriberg *et al.* (1996) study, and therefore were not included in this study. A description of these will not be mentioned.

- Group 1: Sentences containing two accented noun phrases

The first group involves two noun phrases with a total of four accented words; there are two subtypes, one in which both noun phrases have two accented words ("2-2") and one in which the first has three and the second only one ("3-1"). These were paired lexically, as shown in the following examples. Accented words are written in capital letters:

2-2: Je moet de MOOIE ROZEN in een GELE VAAS doen

(You should put the pretty roses in a yellow vase)

3-1: Je moet de MOOIE GELE ROZEN in een VAAS doen

(You should put the pretty yellow roses in a vase)

2-2: We hebben de LELIJKE LAKENS op de OUDE SOFA gelegd

(We put the ugly sheets on the old sofa)

3-1: We hebben de LELIJKE OUDE LAKENS op een SOFA gelegd

(We put the ugly old sheets on a sofa)

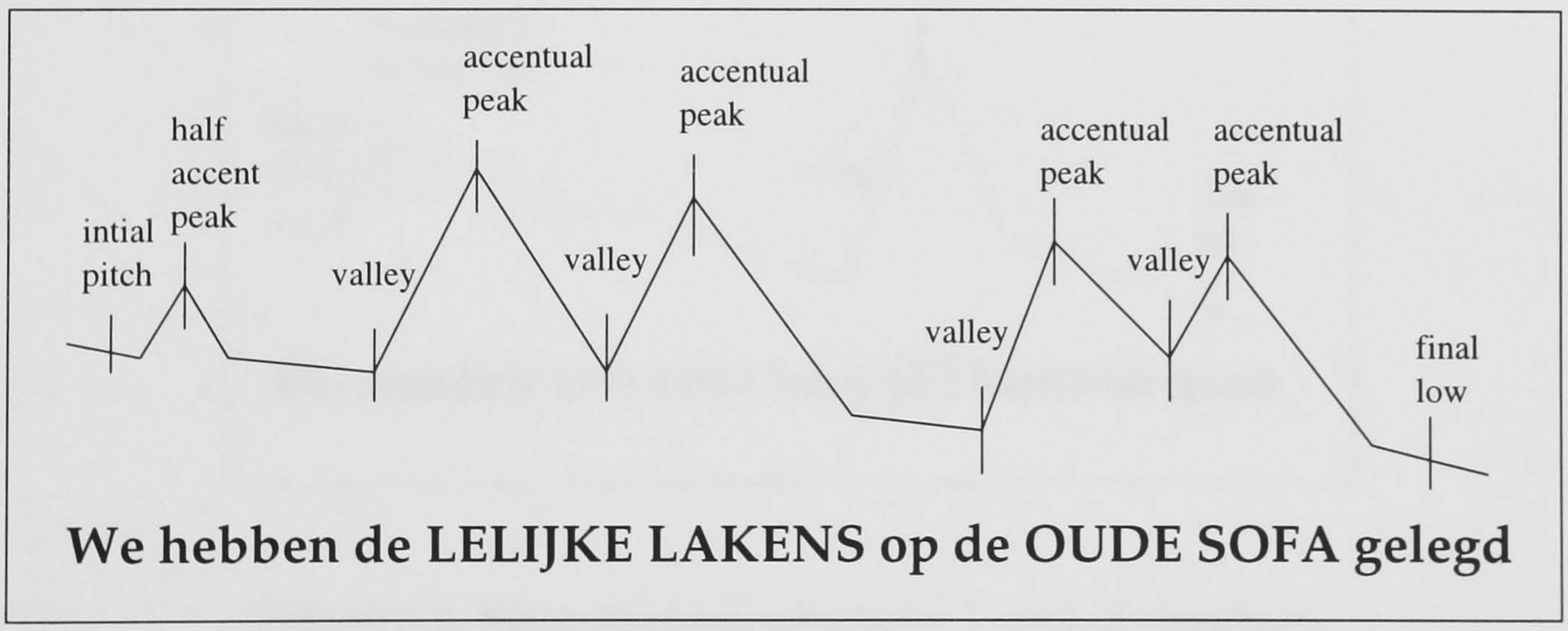



**Figure 3.1.** Measurement targets for Group 1 sentences

Altogether there were eight lexical pairs, and each sentence was read twice, for a total of 32 utterances in group 1. From the 32, seven of the 2-2 utterances and seven of the 3-1 utterances were selected for inclusion in our own experiment. The choice of

utterances that were included/discarded was such that there was as much variation in the utterances presented. The criterion as to the number of utterances choosen was based on getting approximately a minute's worth of speech for each speaker.

Target levels that were studied in this group were initial pitch, half accent peaks, accentual peaks, medial valley between the two noun phrases, valleys immediately preceding accents and the final low. These levels are represented schematically in diagram 3.1.

- Group 2: Long single-accent sentences

These were all of the form

We zouden wel eens naar [X] kunnen gaan
(We really ought to be able to go to [X] sometime),

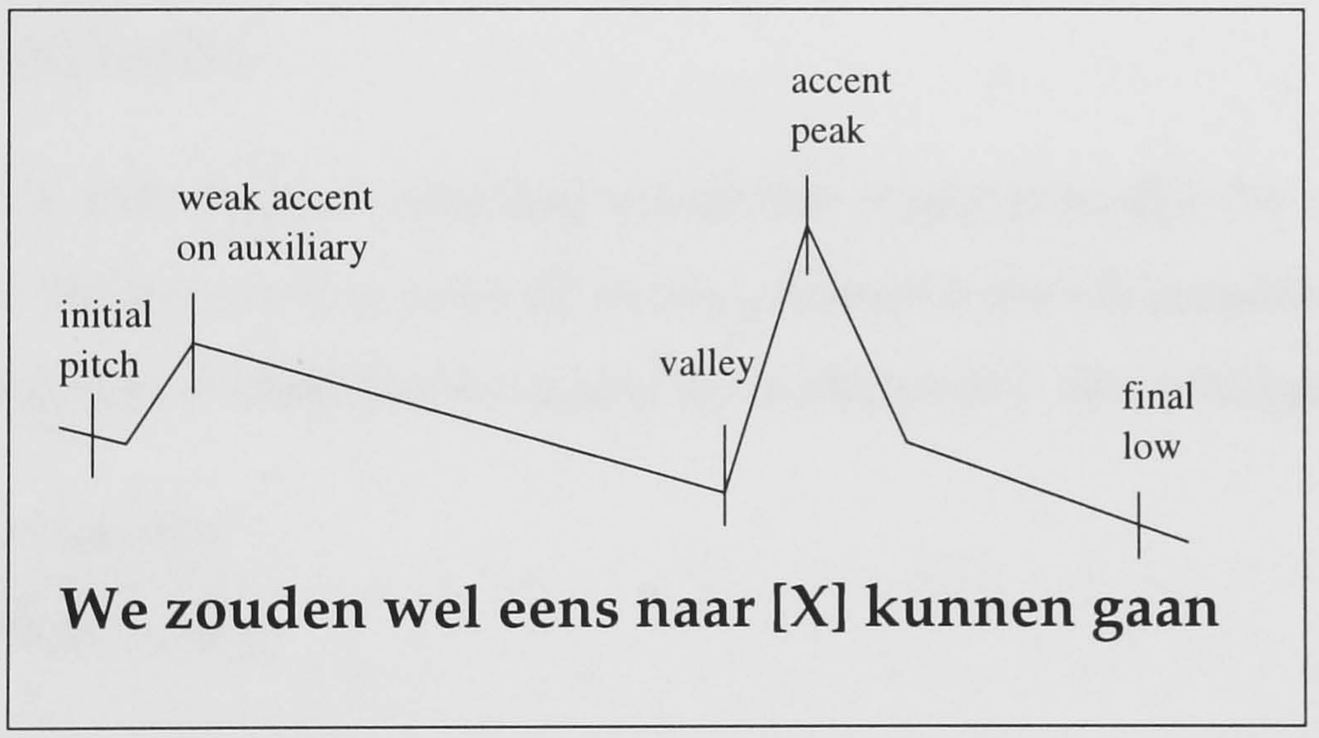



**Figure 3.2.** Measurement targets for Group 2 sentences

in which X was one of four places (Londen (London), Malta, Wenen (Vienna), and Miami) where one might plausibly go on holiday. Each of the four versions was read four times, for a total of 16 utterances. These sentences were designed so that there should be an accent only on the place name, but many speakers put a weak accent

on the auxiliary "zouden" as well. From the 16, three of these utterances were included in our own experiment. From the three that were selected, all the cities except Miami were included. It was felt that the word Miami stuck out much too clearly as an English word, even though spoken by Dutch speakers, and it was important that nothing should be understood by the listeners in the perception experiment which will be detailed subsequently.

Target levels that were studied in this group were initial pitch, weak accent on auxiliary, valley immediately preceding accent, accent peak, and final low. These levels are represented schematically in diagram 3.2.

- Group 3: Contrast sentences

These were all of the form

Ik zei niet [X], maar [Y]
(I didn't say [X], but [Y],

where X and Y were similar-sounding words that might plausibly be confused in a real situation. There were four pairs of words, presented in both possible orders, with two repetitions of each sentence, for a total of 16 utterances. The word pairs were

mannetjes / lammetjes
(little men / little lambs)

rommelen / morrelen
(fiddle / tinker)

Malika / Monica
(women's names)

namelijk / mannelijk
(namely / masculine)

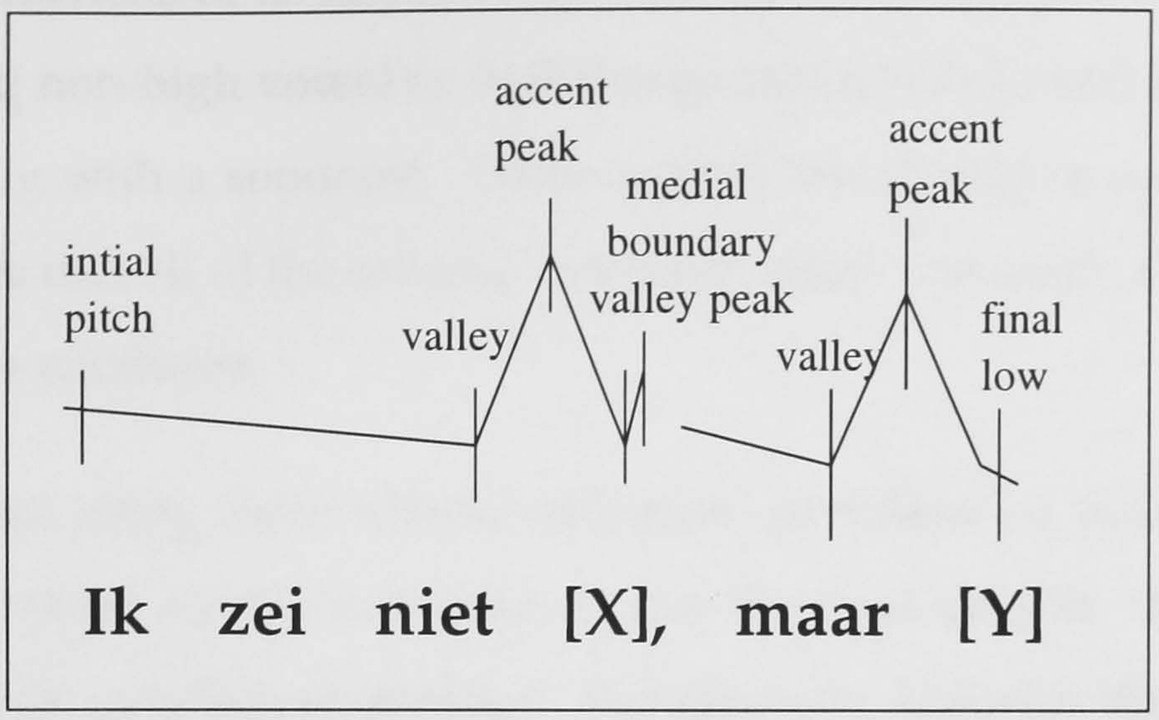



**Figure 3.3.** Measurement targets for Group 3 sentences

Again three utterances were selected from the 16 available and for a similar reason to that given for the selection of the group 2 utterances, the Malika/Monica utterances were discarded due to the recognisable name to English listeners.

Target levels that were studied in this group were initial pitch, valleys preceding accents, accent peaks, final low and both valley and peak of medial boundary rise. These levels are represented schematically in diagram 3.3.

### Criteria used in constructing sentences

In constructing the sentences for the Shriberg *et al.* (1996) study, prosodic, pragmatic and segmental phonetic considerations were taken into account. Prosodically, all the sentences of a group had similar rhythmic patterns, and contained no sequences of accented syllables without at least one intervening unaccented syllable. Pragmatically, the sentences were intended to be reasonably natural sentences that might actually be spoken in a real situation, although clearly some are more awkward than others. After these considerations, if there was at all a possible choice, words with high vowels and obstruents were avoided. This was to minimise intrinsic f0 effects and to minimise

segmental perturbations of f0 respectively. The ideal test syllable contained a sonorant onset, a long non-high vowel or diphthong, and was followed by an unstressed syllable beginning with a sonorant. These criteria are all met in e.g. "lelijke". Not every test syllable met all of the criteria, but every effort was made to keep violations of the criteria to a minimum.

For utterance final pitch, there was by definition no following syllable, but insofar as possible, utterances were constructed so that the final syllable otherwise met the segmental phonetic criteria just sketched. Furthermore, sentence-final syllables were intended to be (a) unaccented, but if possible (as in all the sentences ending with infinitives or past participles) lexically stressed, and (b) separated from the last accented syllable by at least one intervening syllable. None of the sentence-final syllables met all the segmental phonetic criteria, but all were unaccented, and all but one were separated from the last accented syllable by an intervening syllable. Utterances in which the speaker accented the final infinitive or past participle were excluded from the analysis.

### 3.2.3 Recordings

Each speaker recorded all the materials in a single recording session lasting about 75 minutes, with a short break halfway through. The recordings were made in a quiet recording studio at IPO using professional equipment. Speakers were seated comfortably in front of a table with a computer terminal on it. A microphone was placed on either side of the speaker, each approx 10 cm from the speaker's mouth. Recordings were made onto digital audio tape.

The sentences to be read were presented one at at a time on a computer screen placed on a table in front of the speaker. The experimenter, seated in a neighbouring control room, controlled the presentation: when the speaker had finished one sentence satisfactorily, the experimenter pressed a key which caused the next sentence to be presented after a two second delay. This arrangement made it possible to allow extra

time between sentences if, for example, the speaker stumbled and repeated a sentence. None of the speakers complained that the presentation was too slow or two fast.

Taped instructions were presented over loudspeakers in the recording studio. There was a set of general instructions at the beginning, and specific instructions for each section. Taped instructions rather than spontaneously spoken instructions were used to insure consistency across recording sessions.

Speakers were instructed to read each sentence in a relaxed, natural way, and to read each as if it were a separate utterance. No special instructions were given about the intonation to be used for the sentences that have been used for this study.

### 3.2.4 Pitch Range Analysis

The DAT recordings were transferred to the computer system at IPO and separate speech files were made for each sentence.

Extraction of f0 was done by means of GIPOS, an interactive wave form processing package developed at IPO. On the basis of simultaneous time-aligned displays of the waveform and the f0 trace, a number of points were selected as the representative f0 values for each sentence. These points were intended to be relatively stable and reliably identifiable points in the contour; most of them were clear peaks and valleys in the f0 trace.

More specifically, the criteria for selecting measurement points were as follows:

1. for the high pitch of accented syllables, if there was a clear local f0 peak, the f0 peak was chosen, irrespective of its precise alignment with the waveform. In general, as is normally the case in English, these f0 peaks were aligned later than the energy peak of the accented syllable, and in some cases they were aligned early in the following syllable.

2. for the low pitch of unaccented words except at the end of the utterance, a clear local f0 minimum was chosen if there was one (but cf. criterion 4 below).

3. for both high and low points, if there was no local f0 maximum or minimum, the pitch was measured at the energy peak of the accented syllable (for highs) and the specified unaccented syllable (for lows); the energy peak was estimated by eye from the waveform display.

4. for both high and low points, an effort was made to avoid choosing values that appeared to be due to segmental perturbations of f0 (also called "microprosody"; cf. Silverman 1986, 1987). These arose, for example, in the case of the low f0 point for the unaccented word "maar". Whenever the speaker used a strongly articulated velar or uvular /r/, there would be a very local dip aligned roughly at the end of the long /a/ vowel. In cases like these the last value before the beginning of the microprosodic dip was chosen as the f0 value.

5. Finally, for utterance final lows, an effort was made to choose the lowest f0 value that appeared to be reliable. For utterances ending with long vowels and/or sonorants (e.g. all those ending with the word "gaan"), this generally meant taking an f0 minimum as much as 100 ms before the end of phonation, since such minima were often followed by a very slight increase in f0 (this phenomena is observed in other languages as well). For utterances ending with obstruents, the f0 normally dropped fairly rapidly until the end of phonation, and in these cases the last extracted value was chosen unless it was obviously part of a microprosodic dip, in which case a slightly earlier value was taken. For certain speakers, the final one or two syllables were often so irregularly voiced that the pitch extraction failed, and in these cases the last reliable extracted value was taken. In a few cases there were octave errors at the final lows, which were corrected by editing the data files.

These procedures represent a compromise between precision and practicality. It would be possible to obtain more accurate f0 values for individual data points by measuring

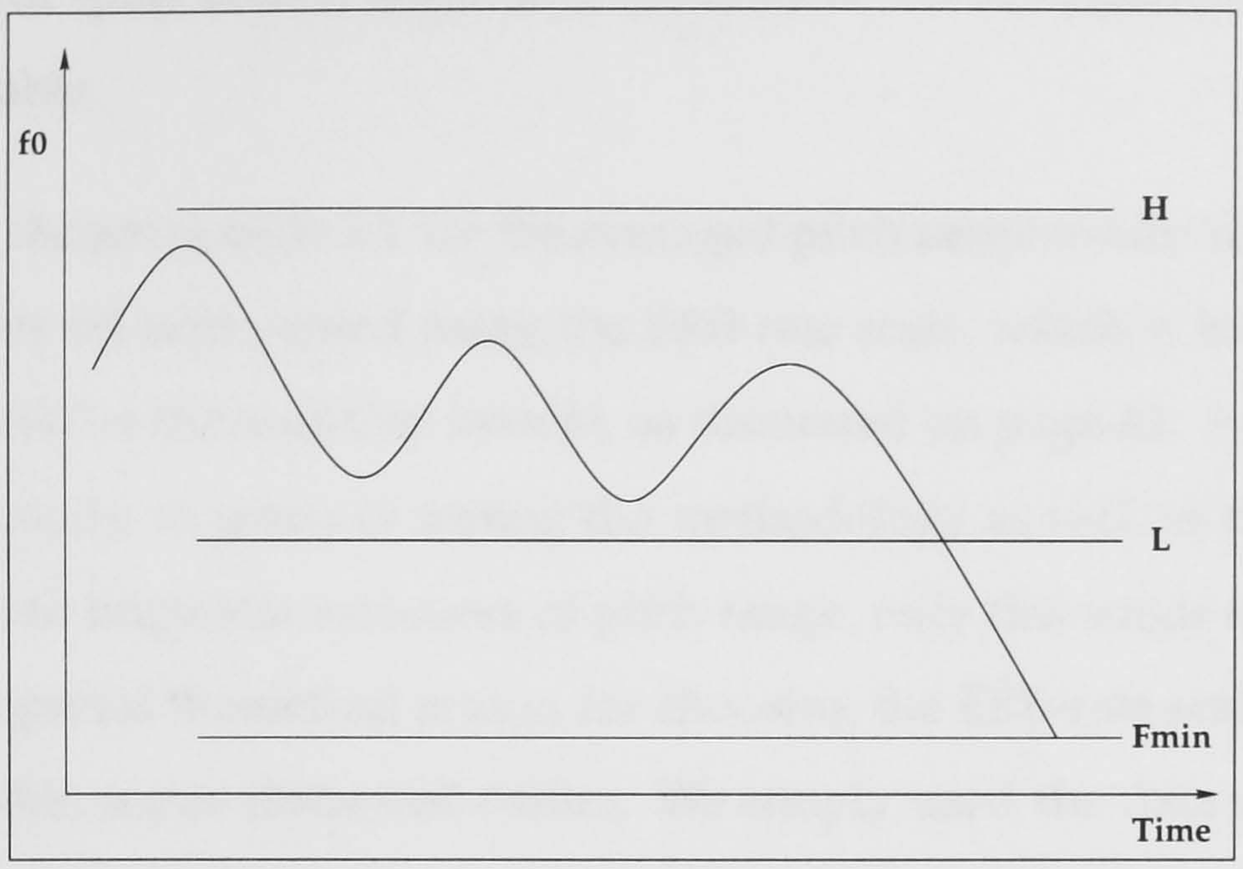



**Figure 3.4.** Pitch Range Variables

the duration of pitch periods, especially in the case of utterance-final lows. However, the time required for such analyses would have made it impossible to undertake the large-scale comparison across speakers and conditions that those involved in the Shriberg *et al.* (1996) study had in mind. They chose to compensate for the measurement error that their procedures undoubtedly introduced into the data by using a relatively large corpus of utterances - most of the mean values used in the current study were based on measurements of 16 utterances, even though we only used a small subset of the recordings for our own purposes.

## 3.3 Pitch Range Results

Figure 3.4 shows a schematic representation of an idealised f0 contour with a visual description of the measurement points that were used for modelling speakers' pitch range. It should be clear from this figure that "H" indicates a potential topline of the speaker span, "L" indicates a potential bottomline of the speaker span and "Fmin" indicates the sentence-final low which will also be considered to be a potential bottomline for span as well as a potential measure for level. No other measures of topline

or bottomline for span, and no other measure of level will be considered as no further data was available.

The set of data shown in table 3.1 are the averaged pitch range results for each speaker. These results are all represented using the ERB-rate scale, which is based on the frequency selectivity of the auditory system, as discussed on page 42. Again, given the nature of this study, in terms of testing the methodology as well as the preliminary investigation into linguistic measures of pitch range, only this single scale was used. There was no special theoretical reason for choosing the ERB-rate scale compared to the other possible scales discussed earlier. We simply used the data as it was given in the Shriberg *et al.* (1996) study. In table 3.1 it can be seen that there is a reasonable spread of results. Speaker JR has the lowest level and Speaker EV has the highest level. Speaker LV has the widest span using either measure and the narrowest span using the H-L measure is speaker RS, while the H-Fmin measure shows that speaker EV has the narrowest span. As is clear, in this experiment only two different measures of span are going to be assessed, and already slight differences in how they are measured means it is not clear whether RS or EV has the narrowest span. It is this difference in the nature of the span measure which is one of the main investigations being undertaken. It is the aim of this thesis to clarify the nature of pitch range and how it should be measured.

Figure 3.5 shows some of the information from table 3.1, giving a visual representation of the span and level measurements for all the eleven speakers in a scattergraph. For the purposes of this figure, the H-L span measure was used. From this figure it can be seen that level and span measures do seem to be independent with there clearly being speakers that have a narrow span yet with a spread of differing levels. Likewise there are speakers that have very similar levels with a wide range of spans. Given the fact that men and women are readily distinguishable by the level of their voices, it is not surprising that there is a clustering of all the male speakers at the lower end of the y-axis (representing level) in the figure, with the female speakers clustering at the top end of the y-axis.

| | | Span | | Level |
|---|---|---|---|---|
| Speaker | Sex | H-L (ERB) | H-Fmin (ERB) | Fmin (ERB) |
| AC | female | 1.14 | 1.75 | 5.12 |
| ES | female | 1.26 | 1.86 | 5.09 |
| EV | female | 1.27 | 1.43 | 5.57 |
| IS | female | 1.54 | 2.15 | 5.12 |
| JR | male | 1.25 | 1.79 | 2.32 |
| LV | female | 2.19 | 2.71 | 4.78 |
| MH | male | 1.00 | 1.76 | 3.54 |
| RE | male | 1.38 | 1.87 | 3.02 |
| RS | male | 0.90 | 1.79 | 2.73 |
| RW | male | 1.75 | 2.39 | 3.14 |
| UA | female | 0.95 | 2.31 | 4.39 |

**Table 3.1.** Data taken from Pitch Range Modelling Experiment

## 3.4 Perception Experiment

The aim of this perception experiment is to get profiles of speaker characteristics based on a selection of pragmatic and phonetic criteria for each speaker (for example speaker JR is rated as "X" confident and "Y" tense compared to speaker IS who is rated as "A" confident and "B" tense). This was achieved by asking subjects to fill in a rating form while listening to speech recordings of each speaker. From the listeners' responses it is possible to see how each speaker can be characterised. With this data it is then possible to compare the perception of speaker characteristics with the data showing across-speaker variation in pitch range. Therein lies the central point of the experiment. It is these correlations that will be used to firstly support the previous findings, showing a consistent variation in pitch range with variation in speaker characteristics. Further on in the thesis it is the strength of these correlations between pitch range variables and listener's judgements of speaker characteristics which will be used to assess the various suggested measures of range as described in chapter 2.

Given the subjective nature of the task, it is generally acknowledged that listeners are surprisingly reliable on how they judge speaker characteristics, both within- and

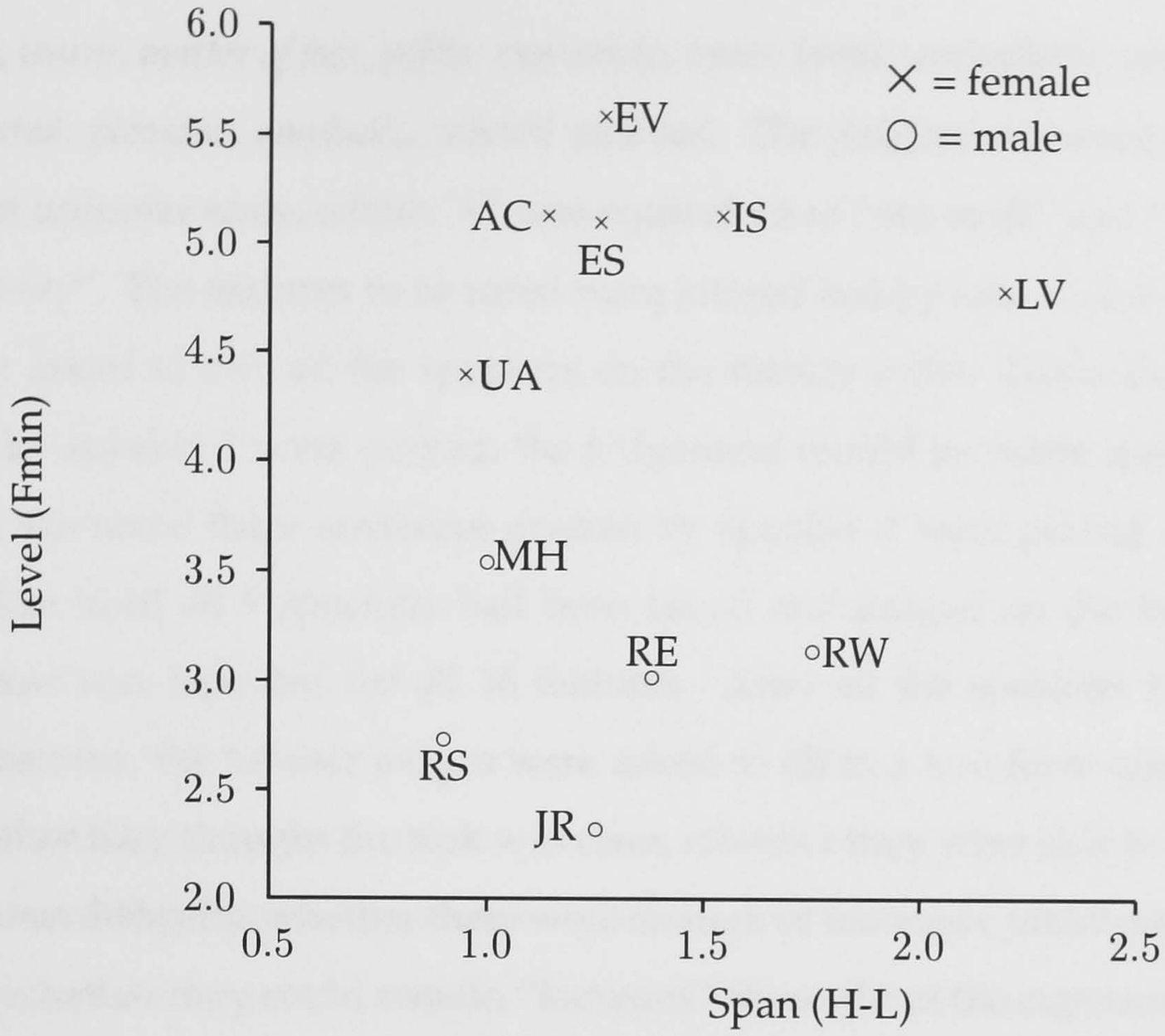



**Figure 3.5.** Span and Level of the 11 Dutch Speakers

across-listener (Brown *et al.* 1974). Also as there is a long history in this type of work, the types of features for which speakers can be reliably judged are well established (Bezooijen 1984, Scherer 1988). It is important to establish a method that will allow for the collection of valid results but also to establish a routine that will allow for a combination of many speakers being judged on as many characteristics as possible.

### 3.4.1 Perception Study: Pilot work

Two experimental design variations were tested before deciding on the most appropriate design. We will briefly describe both methods.

**Pilot study: Method 1**

A panel of 6 English subjects were asked to rate 9 speakers of Standard Dutch, 5 female and 4 male, on 16 phonetic and pragmatic criteria. The features judged were

*deep, warm, matter of fact, polite, expressive, tense, bored, unemphatic, weak, nasal, confident, irritated, pleasant, emphatic, relaxed* and *sad*. The judgements were made on a seven-point unipolar scale, where "1" was equivalent to "not at all" and "7" was equivalent to "very". The features to be rated were judged one by one. So for example, listeners were asked to rate all the speakers on the feature *polite*. Three short sentences spoken by speaker 1 were played, the judgement would be made during a brief pause, then the same three sentences spoken by speaker 2 were played continuing in this fashion until all 9 speakers had been heard and judged on the feature *polite*. This process was repeated for all 16 features. After all the speakers had been rated on all features, the listener judges were asked to fill in a free-form questionnaire asking whether they thought the task was clear, whether they were able to complete the task without difficulty, whether there were enough or too many utterances by each speaker and whether they could remain "focussed" throughout the experiment.

This method involves a strict presentation scheme but is very time consuming. The experimental session took 45 minutes. Only few utterances per speaker were used in this method and there were no guarantees that the utterances selected bore much resemblance to the pitch range data for each speaker which was averaged over a much greater number of utterances.

**Pilot study: Method 2**

A different panel of 6 subjects were asked to rate the same 9 speakers of Standard Dutch on the same 16 phonetic and pragmatic criteria as used in the first method described directly above. The judgements were also to be made on the same seven-point unipolar scale. This time 12 utterances all by speaker 1 were played in quick succession and listeners were asked to make their judgements straight away on the first 8 features while the speech was still playing. A brief pause of five seconds was allowed for the listeners to finish off making their decisions, then the same 12 utterances were repeated and judgements on the next 8 features were carried out. After another brief

pause of 5 seconds, the same procedure was carried out for the next speaker. After all 9 speakers had been judged by this method the same questions as in the first pilot study (page 73 above) were answered by the listener judges.

This method does not allow as much experimental control, that is the experimenter could not be sure which utterance was having the most influence over the listeners when they were giving their responses to the features. This method did offer a much quicker presentation method though. The experimental session took just over 10 minutes. Also more utterances were used so these were likely to bear closer resemblance to the averaged pitch range data for each speaker.

**Pilot Study: Results**

| | | | | |
|---|---|---|---|---|
| confident | $\rho = 0.440$ | $p < 0.001$ | one tailed test | N = 54 |
| deep | $\rho = 0.517$ | $p < 0.0001$ | one tailed test | N = 54 |
| expressive | $\rho = 0.459$ | $p < 0.0005$ | one tailed test | N = 54 |
| irritated | $\rho = 0.316$ | $p < 0.05$ | one tailed test | N = 54 |
| nasal | $\rho = 0.372$ | $p < 0.01$ | one tailed test | N = 54 |
| sad | $\rho = 0.414$ | $p < 0.005$ | one tailed test | N = 54 |
| bored | $\rho = 0.266$ | $p > 0.05$ | one tailed test | N = 54 |
| emphatic | $\rho = 0.175$ | $p > 0.1$ | one tailed test | N = 54 |
| pleasant | $\rho = 0.193$ | $p > 0.1$ | one tailed test | N = 54 |
| polite | $\rho = 0.175$ | $p > 0.1$ | one tailed test | N = 54 |
| tense | $\rho = 0.251$ | $p > 0.05$ | one tailed test | N = 54 |
| unemphatic | $\rho = 0.245$ | $p > 0.05$ | one tailed test | N = 54 |
| warm | $\rho = 0.257$ | $p > 0.05$ | one tailed test | N = 54 |
| weak | $\rho = 0.187$ | $p > 0.1$ | one tailed test | N = 54 |
| matter of fact | $\rho = 0.077$ | $p > 0.5$ | one tailed test | N = 54 |
| relaxed | $\rho = 0.034$ | $p > 0.8$ | one tailed test | N = 54 |

**Table 3.2.** Results of correlation analyses on pilot study data

The results for the two experiments were similar for most of the features being investigated. The results for each feature between the two different experimental conditions (as shown in table 3.2) all showed positive correlations. *Deep, nasal, expressive, confident, irritated* and *sad* showed significant correlations ($p < 0.05$) across the two

experimental conditions. *Weak, pleasant, warm, tense, polite, emphatic, bored* and *un-emphatic* showed correlations that didn't quite reach a suitable level of significance, though considering the degrees of freedom and the subjective nature of the task this is still a satisfactory result. Only *matter of fact* and *relaxed* failed to show much correlation between the two experimental conditions. As method 2 proved to be easier for the listeners to do and involved a lot less time, this style was used for the main experiment.

### 3.4.2 Speech Materials

The full complement of 11 speakers was used, 6 female (*AC, ES, EV, IS, LV, UA*) and 5 male (*JR, MH, RE, RS, RW*). The speech materials used for this part of the experiment were a subset of the utterances used in the Shriberg *et al.* (1996) study. Four different types of sentences were used namely the "3-1" type, the "2-2" type, the "holiday" type and the "contrast" type (cf. section 3.2.2).

Two experimental tapes were prepared. On each tape was all the speech material for a full experimental run. This consisted of three separate presentations of the speech of the 11 speakers as well as a single presentation of one speaker that could be used for a trial run. This trial run was used so the listener judges had the opportunity to become familiar with the task that was being asked of them. The stimuli were played to groups of subjects on a high quality tape machine with high quality speakers in a large room.

Each presentation of a speaker consisted of 16 sentences. These 16 sentences were made up of 5 X the "3-1" type, 5 X the "2-2" type, 3 X the "holiday" type and 3 X the "contrast" type. No two sentences of the same type were adjacent on the test tape. The two experimental tapes used exactly the same speech materials, the only difference being that the speakers were presented in a different order. As there were 11 speakers, 6 female and 5 male, each presentation started with a female voice, which was followed by a male voice. This alternation carried through all of the 11 speakers.

### 3.4.3 Listener Judges

There were 30 subjects in 4 experimental sessions. The majority of subjects were linguistics students at varying levels, from first year undergraduate level to PhD students near completion of their work. All subjects were native speakers of English who did not know Dutch. 15 subjects listened to the first experimental tape while the remaining 15 listened to the second. A brief check of the listener judges' responses was made to make sure there were no obvious "erroneous" marking schemes. Examples of listeners not completing the task correctly would be either the same number being circled for all features for all voices, or perhaps every number circled for all features and for all voices. Only one such response sheet was found, in which the subject decided to circle every number from 1 to 7 for every feature and every speaker. These responses were discarded and the data for 29 subjects has been used for analysis.

### 3.4.4 Rating Forms

The experiment was broken up into three sections, each with its own rating form. In all, there are twenty features that were to be judged, but dividing the experiment into three sections allowed the duplication of ten of the features (*deep, expressive, pleasant, nasal, creaky, bored, relaxed, breathy, sad* and *irritated*). This overlap was intended as a check on the consistency of the raters. The features that speakers were being judged on in each rating form are shown in table 3.3. The same seven point unipolar scale used in the pilot studies was used for this experiment; see section 3.4.1.

The features selected for the current experiment were based on a number of criteria. While our main interest is to focus on criteria that are expected to be affected by pitch range variation, we are also interested to see if our data will pattern in a similar way to previous research (Uldall 1964, Brown *et al.* 1973). For the current experiment there are a number of features that reflect the nature of a speaker's character (e.g. *expressive, pleasant, bored, relaxed, emphatic, confident*). There are also a number of features which

relate to voice quality (e.g. *whisper, creaky, nasal, tense*). Also there are some control items put in to make sure that the listener judges are being consistent and making sensible judgements. Clearly listener judges should not be characterising speakers as being both *emphatic* and *unemphatic*, both *happy* and *sad*.

All the features selected have come from features used in a long line of previous research of a similar nature including Uldall (1964), Huttar (1968), Scherer *et al.* (1984), Bezooijen (1984), and Ladd *et al.* (1985). Based on this previous research, initially discussed in section 2.2, we expect that the more positive attributes (*confident, happy, expressive, emphatic, pleasant*) will correlate positively with pitch span while the more negative attributes (*weak, irritated, sad, unemphatic*) will correlate negatively with pitch span. We would predict that *relaxed, emphatic* and *unemphatic* would correlate with level. *Relaxed* and *unemphatic* would correlate negatively with level due to being reflective of low arousal, while *emphatic*, being high arousal, would correlate positively with level. These predictions are following the results of Ladd *et al.* (1985). *Deep* is a feature that should strongly correlate with level. This is essentially a control feature for level. There would be a clear flaw in the methodology if listener judges considered speakers to have a deep voice if our potential measures of level marked the speakers as having a high voice.

The main reason for including voice quality features at this stage is to follow the patterns of previous research (Uldall 1964, Brown *et al.* 1973), and to see if there are any interesting correlations of these voice quality features with the other features that we are considering to be strongly related to pitch range. It is not expected that features of voice quality will have any correlations with the pitch range variables and in that sense could also be considered good control items for testing the validity of our methodology.

| Features | | |
|---|---|---|
| Rating form 1 | Rating form 2 | Rating form 3 |
| deep | emphatic | polite |
| expressive | matter of fact | unemphatic |
| weak | bored | irritated |
| pleasant | confident | sad |
| warm | relaxed | breathy |
| whisper | creaky | harsh |
| nasal | tense | nasal |
| creaky | breathy | pleasant |
| bored | sad | deep |
| relaxed | irritated | expressive |

**Table 3.3.** Features used in the three rating forms

### 3.4.5 Experimental Session

Subjects were given three rating form booklets. On the cover of the first rating booklet, there were general instructions as to how the experiment would be run, as well as an example of the rating form. Once they had finished reading the instructions, the subjects had an opportunity to ask further questions. Then a practice run was carried out, using the speech of one speaker only. After a further opportunity for questions, the experiment proper was run. Each of the three sections took around 12 minutes to complete. After each section was completed, the rating form booklets for that section were collected in by the experimenter. This was to prevent any subject who had noticed that there was a pattern in the ordering of speakers and noticed that some of the features were duplicated from checking to see what he/she had put for a previous answer. The whole experimental session lasted for 45 minutes.

## 3.5 Results

The *mode* and *median* for each feature for each speaker averaging across all listeners were calculated using the SPSS statistical package. The mode data, which was primarily used for further analyses, can be found in table 3.4. In table 3.4 it shows, for example, that speaker AC was judged as being a "2" on the *deep* scale and "5" on the *relaxed* scale. So according to the modal score, generally the listeners perceived speaker AC as not having a very deep voice, but sounding reasonably relaxed. For the same variables *deep* and *relaxed*, speaker JR was judged by the majority of listeners as being "7" and "6" which means that JR is perceived as having a very low voice and sounds very relaxed. The median data was used as input for the multi-dimensional scaling discussed on page 81. There was very little difference between the two averaging techniques. Although there are arguments for using any of the averaging techniques (mode, median or mean), given the nature of the data collected, the mode is the most robust (Hatch & Lazaraton 1991). Although there was data collected for features that were included within the three rating forms twice, this data has not been included in table 3.4.

Spearman's rank correlation coefficients ($\rho$) were calculated for each pair of features. Table 3.5 briefly summarises those features that correlate significantly (at least $p < 0.05$). These results are unsurprising, for example the fact that *expressive, pleasant* and *warm* all show significant positive correlations with each other. The predictability of these results helps verify that the methodology is valid. A brief summary of table 3.5 shows that the features *pleasant, relaxed, polite, warm* and *confident* are positively correlated. *Weak, whisper* and *sad* are also positively correlated. These two sets of features (the *"pleasant"* set vs. the *"weak"* set) are negatively correlated i.e. the higher the scores for the first group the lower the scores of the second group.

For the next stage we attempted to establish relationships between the pitch range parameters for each speaker (table 3.1) and the results of the judgement study (table 3.4)

| | Speakers | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | AC | ES | EV | IS | JR | LV | MH | RE | RS | RW | UA |
| deep | 2 | 1 | 1 | 1 | 7 | 1 | 4 | 5 | 5 | 6 | 2 |
| expressive | 3 | 5 | 3 | 5 | 5 | 5 | 2 | 4 | 2 | 5 | 3 |
| weak | 2 | 5 | 5 | 2 | 1 | 2 | 5 | 2 | 4 | 1 | 3 |
| pleasant | 5 | 4 | 4 | 5 | 5 | 5 | 2 | 4 | 4 | 4 | 4 |
| warm | 5 | 4 | 3 | 5 | 5 | 5 | 3 | 4 | 3 | 4 | 5 |
| whisper | 1 | 3 | 3 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 |
| nasal | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |
| creaky | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| bored | 3 | 3 | 2 | 1 | 2 | 3 | 4 | 3 | 3 | 2 | 4 |
| relaxed | 5 | 5 | 4 | 5 | 6 | 5 | 3 | 5 | 5 | 5 | 5 |
| polite | 5 | 5 | 5 | 5 | 4 | 5 | 3 | 5 | 5 | 4 | 5 |
| unemphatic | 3 | 3 | 5 | 3 | 3 | 3 | 6 | 2 | 4 | 2 | 4 |
| irritated | 2 | 2 | 1 | 1 | 2 | 3 | 3 | 3 | 2 | 3 | 2 |
| sad | 2 | 4 | 5 | 1 | 2 | 1 | 2 | 1 | 3 | 1 | 2 |
| breathy | 3 | 5 | 3 | 3 | 2 | 2 | 3 | 1 | 2 | 2 | 3 |
| harsh | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| emphatic | 2 | 3 | 3 | 4 | 4 | 5 | 2 | 3 | 3 | 5 | 2 |
| matteroffact | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 3 | 5 |
| confident | 5 | 5 | 5 | 5 | 6 | 5 | 4 | 5 | 5 | 6 | 5 |
| tense | 5 | 2 | 2 | 2 | 1 | 1 | 3 | 3 | 1 | 1 | 3 |

**Table 3.4.** Mode results for all speakers

by calculating Spearman's rank correlation coefficients ($\rho$). Table 3.6 shows which features significantly correlated ($p < 0.05$) with these measures of range. Five features - *deep, nasal, sad, harsh* and *breathy* - correlate with the level parameter (Fmin). Seven features - *expressive, bored, unemphatic, sad, emphatic, pleasant* and *creaky* - correlate with the span measure (H-L). Of these 7 features, 5 of them also show strong correlations with the alternative span measure (H-Fmin). As can clearly be seen in table 3.6 by the coefficients marked with a tick, in all but one of these instances (for the feature *sad*), the H-L shows the strongest correlations with the results of the judgement study. It seems that higher voices are judged as being more breathy and sad, while lower voices are judged as being more deep, nasal and harsh. A wide span correlates with more expressive, emphatic and pleasant judgements while a narrower span is correlated with more bored, unemphatic, sad, and creaky judgements.

| FEATURE | FEATURE | COEFFICIENT ($\rho$) | FEATURE | FEATURE | COEFFICIENT ($\rho$) |
|---|---|---|---|---|---|
| deep | nasal | 0.68 | deep | whisper | -0.55 |
| deep | polite | -0.61 | deep | breathy | -0.64 |
| deep | harsh | 0.87 | expressive | weak | -0.58 |
| expressive | pleasant | 0.55 | expressive | warm | 0.55 |
| expressive | relaxed | 0.58 | expressive | nasal | -0.71 |
| expressive | bored | -0.58 | expressive | unemphatic | -0.72 |
| expressive | confident | 0.62 | expressive | emphatic | 0.78 |
| weak | pleasant | -0.62 | weak | relaxed | -0.72 |
| weak | warm | -0.63 | weak | whisper | 0.70 |
| weak | unemphatic | 0.75 | weak | sad | 0.72 |
| weak | emphatic | -0.59 | weak | confident | -0.76 |
| weak | breathy | 0.62 | pleasant | warm | 0.81 |
| pleasant | relaxed | 0.65 | warm | relaxed | 0.66 |
| whisper | relaxed | -0.52 | whisper | breathy | 0.60 |
| whisper | harsh | -0.59 | whisper | matter of fact | 0.62 |
| nasal | confident | 0.61 | creaky | harsh | 0.55 |
| creaky | matter of fact | -0.64 | creaky | confident | 0.57 |
| bored | confident | -0.69 | bored | emphatic | -0.62 |
| relaxed | unemphatic | -0.61 | relaxed | confident | 0.74 |
| polite | harsh | -0.62 | unemphatic | sad | 0.65 |
| unemphatic | emphatic | -0.56 | unemphatic | confident | -0.61 |
| irritated | breathy | -0.53 | sad | breathy | 0.54 |
| breathy | harsh | -0.54 | emphatic | confident | 0.64 |
| emphatic | matter of fact | -0.59 | emphatic | tense | -0.83 |
| matter of fact | tense | 0.70 | confident | tense | -0.59 |

**Table 3.5.** Significant Correlations of Features

A number of data reduction techniques have been used to condense the huge amount of numbers collected in similar experiments. Given the non-parametric nature of the data collected, the only viable robust option is to use multi-dimensional scaling, which permits the use of median data as opposed to the mode data that was used for the correlation analysis. Again, using the median is a justified averaging technique for rating scale data along with the mode, whereas the more commonly used mean averaging technique is not suitable. The standard use of the data reduction techniques, most notably factor analysis (Uldall 1964, Brown *et al.* 1973), is to aim to reduce the number of features to see which features cluster together. Such an analysis run on the 20 features from the data from the current study confirms the results of the earlier correlation analysis. The features *expressive, emphatic, pleasant* and *warm* cluster in the same area. The features *irritated, tense, harsh* and *creaky* cluster together closely. Also the

| Feature | Range Measurement | | |
|---|---|---|---|
| | Level (Fmin) | Span (H-L) | Span (H-Fmin) |
| deep | -0.876 | | |
| nasal | -0.703 | | |
| sad | 0.613 | | |
| harsh | -0.640 | | |
| breathy | 0.706 | | |
| expressive | | √ 0.763 | 0.588 |
| bored | | √ -0.749 | -0.744 |
| unemphatic | | √ -0.628 | -0.531 |
| sad | | -0.610 | √ -0.710 |
| emphatic | | √ 0.779 | 0.572 |
| pleasant | | 0.593 | |
| creaky | | -0.657 | |

**Table 3.6.** Features that correlate with span and level measurements

reduction process calculated that the features *sad, weak, unemphatic* and *bored* should be near each other in the same region in a 2 dimensional analysis. The results of the multi-dimensional scaling are shown in table 3.7. In this table we have also reported the stress value (0.145) and the squared correlation (RSQ = 0.916). These statistics show that a high proportion of the variation in the data can successfully be accounted for within a two dimensional space. The results shown in table 3.7 are represented graphically in a scattergraph shown in figure 3.6.

Taking a different approach to data reduction, we tried a by-speakers analysis as opposed to a by-features analysis. Considering each feature as a dimension, it is possible to consider a full description of a speaker to be made accurately in a 20 dimensional space. Using multi-dimensional techniques we reduced this description of speakers to a 2 dimensional space. Results of this can be found in figure 3.7a.

This gives an interesting result in that the distribution of speakers in a 2 dimensional space, based on listener judges responses to 20 pragmatic and phonetic criteria produces a remarkably similar figure to that shown in figure 3.5, which is reproduced

| Feature | Dimension 1 | Dimension 2 | Feature | Dimension 1 | Dimension 2 |
|---|---|---|---|---|---|
| deep | 3.1715 | -0.9182 | expressive | 0.0184 | 1.3349 |
| weak | -1.4584 | -1.3427 | pleasant | -0.3796 | 1.0761 |
| warm | -0.0179 | 1.1370 | whisper | -1.2870 | -0.3005 |
| nasal | 0.6926 | -0.4342 | creaky | 0.7332 | -0.1935 |
| bored | -0.4772 | -1.2469 | relaxed | -0.2315 | 0.7676 |
| polite | -0.8185 | 0.9054 | unemphatic | -0.8092 | -0.7296 |
| irritated | 0.3058 | 0.1296 | sad | -1.3208 | -1.6429 |
| breathy | -1.1214 | -0.0085 | harsh | 0.6482 | -0.0491 |
| matter of fact | -0.2813 | 0.0620 | confident | 0.4241 | 0.9023 |
| tense | 0.5695 | 0.0969 | emphatic | 0.6611 | 1.4703 |
| Stress = 0.145, RSQ = 0.91605 | | | | | |

**Table 3.7.** Multi-dimensional scaling results - by features

here as figure 3.7b for ease of comparison between the two figures. The speakers pattern up the scale of both *level* and *1st dimension* axes in a similar order and the speakers pattern across the *span* axis in the opposite way to the *2nd dimension* axis. As a follow up study, we attempted to establish relationships between the pitch range parameters for each speaker with the results of the multi-dimensional scaling procedure. Results of this can be found in table 3.8. The level parameter strongly correlates with the dimension 1 variable ($\rho = 0.907, p < 0.05$, one tailed test, N = 11) and the span parameter correlates strongly with the dimension 2 variable ($\rho = -0.791, p < 0.05$, one tailed test, N = 11).

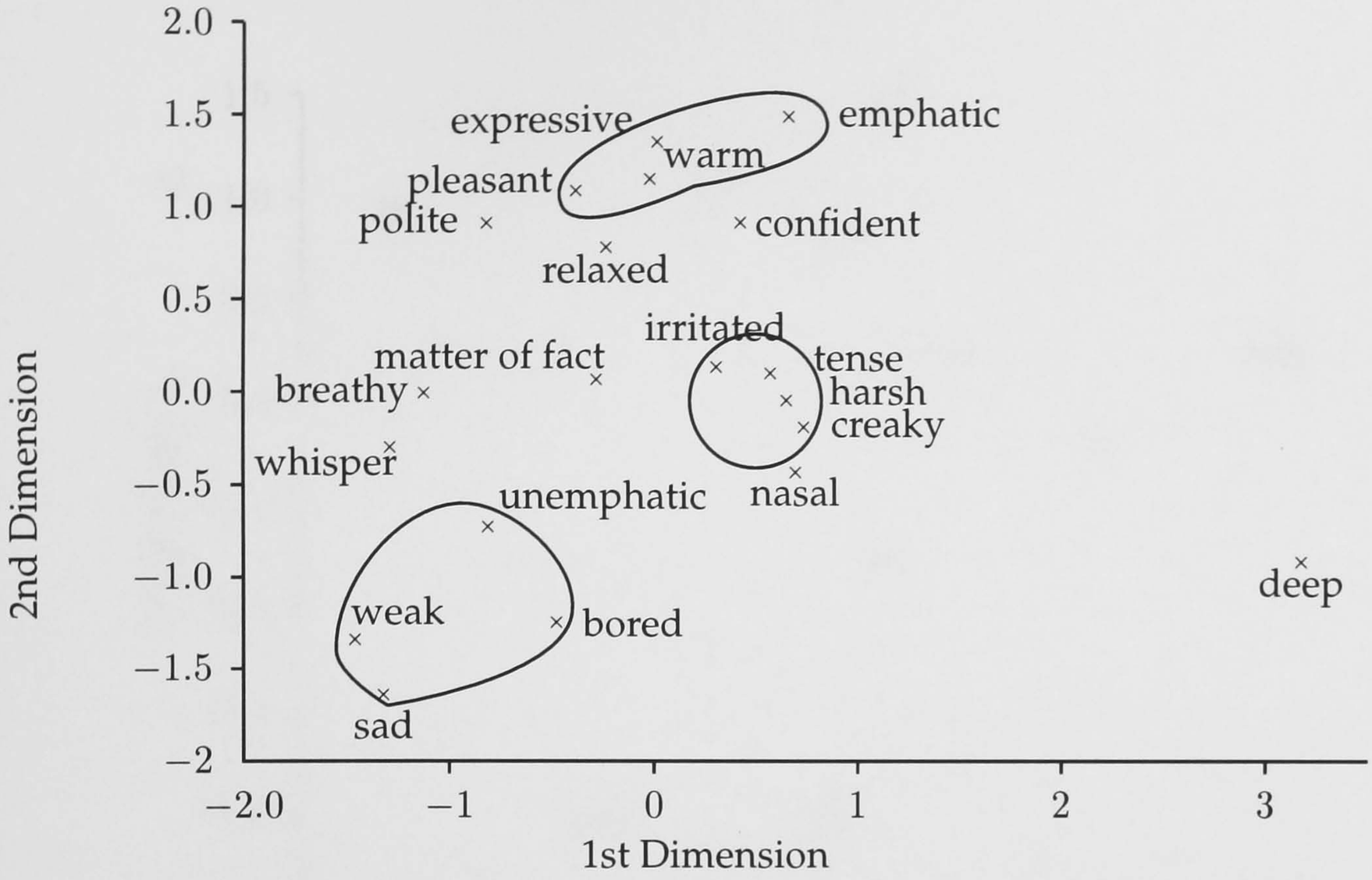



**Figure 3.6.** 20 features characterised in 2D space

| Speaker | Dim.1 | Dim. 2 |
|---|---|---|
| AC | 0.82 | -0.09 |
| ES | 1.08 | -0.28 |
| EV | 1.46 | 0.22 |
| IS | 0.78 | -1.62 |
| JR | -2.00 | -0.04 |
| LV | 0.35 | -1.03 |
| MH | 0.38 | 1.80 |
| RE | -0.65 | 0.29 |
| RS | -0.19 | 1.02 |
| RW | -1.99 | -0.64 |
| UA | -0.27 | 0.37 |
| Stress = 0.092, RSQ = 0.948 | | |
| correlation coefficients $\rho$ | | |
| span (H-L) | -0.410 | -0.791 |
| level (Fmin) | 0.907 | -0.009 |

**Table 3.8.** Multi-dimensional scaling results - by speakers

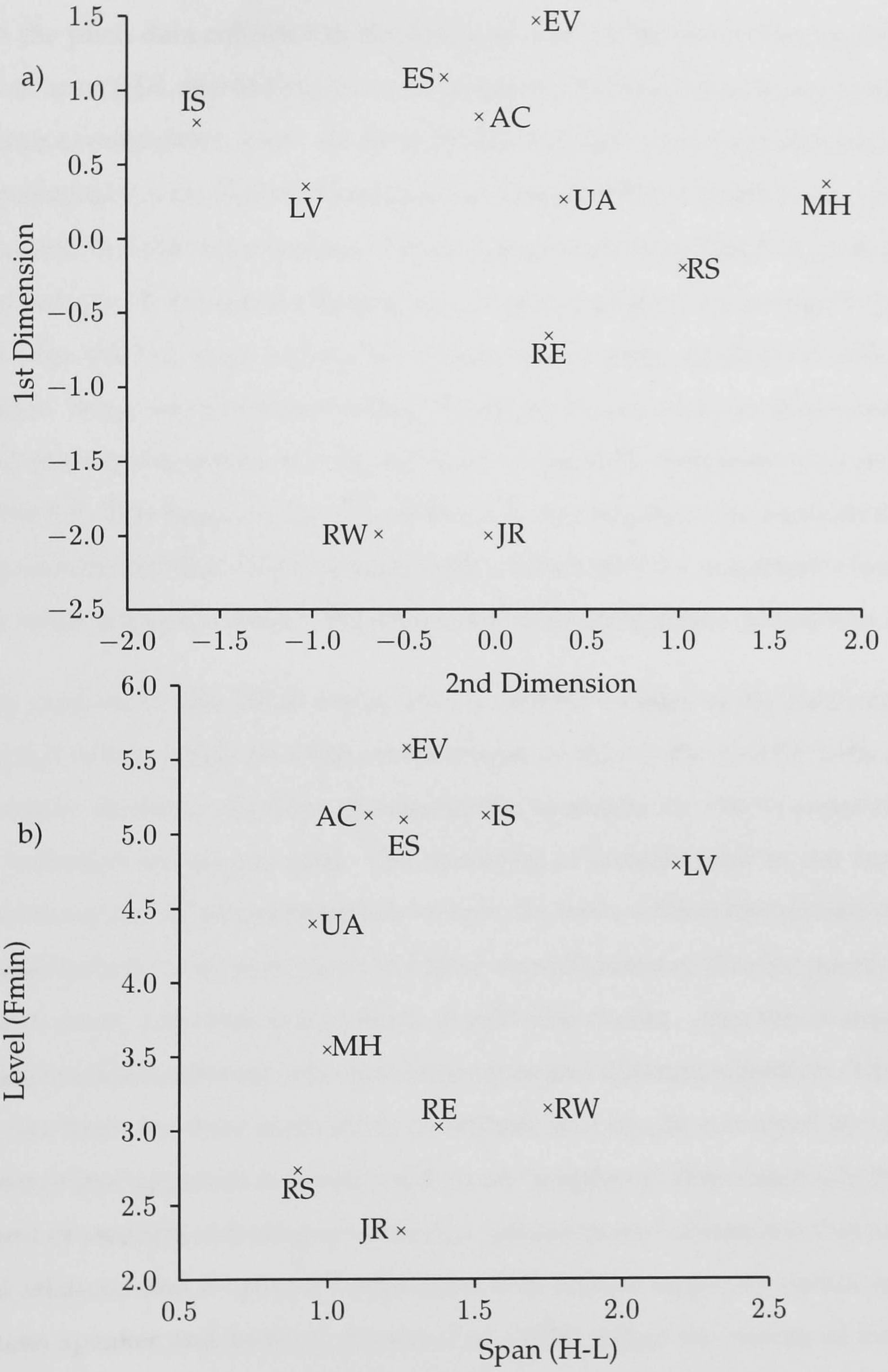



**Figure 3.7.** a) 11 Dutch speakers characterised in 2D space and b) Span and Level of the 11 Dutch Speakers

## 3.6 Conclusions and Discussion

From the pitch data collected in the Shriberg *et al.* (1996) study two possible dimensions of span (H-L and H-Fmin) and one measure of level (Fmin) were correlated with the listener judgement data[2]. All these pitch range data were measured using the ERB psychoacoustic scale used by Hermes & van Gestel (1991). Central to our argument, it is assumed that the best measures of level and span are those which show the strongest correlations with the listener judges' data. For overall span the strongest correlations were with the H-L span and for level there were strong correlations with the Fmin measure. While some of these findings are clearly in line with previous research, what is important is that effects of level and span are partially independent as results show in table 3.6. This supports the hypothesis that two linguistically motivated, partially independent dimensions of variation better characterise the communicative effects of pitch range across speakers, compared to the single dimension of max-min f0.

A key purpose of this Dutch project was to test the validity of the methodology. Although it is impossible to compare the results of this study directly with other representative studies in the field, the similarities in results are clearly supportive of the data collection techniques used. The clustering of features used in our study (as reported in figure 3.6) seems intuitively correct. By this we mean that it must come as no surprise that *sad, weak, unemphatic* and *bored* should cluster or that the positive features *pleasant, warm, expressive* and *emphatic* should also cluster. Also this is supported by studies that used different collection techniques and different adjectives. Uldall (1960) suggests there are three main kinds of attitude that can be conveyed by voice. This 3 dimensional approach is constructed on an "emphasis" dimension which indicates amount or strength of feeling or interest, a "pleasantness" dimension that reflects personal relations and a "power" dimension that reflects authority versus submission between speaker and listener. Brown *et al.* (1973) reflect the results of their feature analysis in a 2 dimensional space that can be based on the dimensions "competence"

[2]These pitch range measures are detailed in section 3.3 and figure 3.4

and "benevolence".

The results from the present study show that the 2nd dimension from the multi-dimensional scaling procedure is very similar to the "competence" dimension of the Brown *et al.* (1973) study or the "pleasantness" dimension of the Uldall (1960) study. Our first dimension seems to be dominated by the voice quality features that we included in our study, which are different to the adjectives used in other studies. This first dimension patterns from *breathy* to *harsh/creaky*. There is an argument that could be made for this dimension patterning in a similar fashion to the "emphasis" dimension. *Polite* and *pleasant* could be considered low feeling positive attributes moving along this "emphasis" dimension through to the high feeling positive attributes *confident* and *emphatic*. The key point of this multi-dimensional scaling is to show that our new experimental techniques continues the same pattern of results with those studies interested in speaker characteristics and emotion in speech.

A suitable measure of span and level can capture some of the differences between speakers, in terms of speaker characteristics, though how much exactly is still unclear. From all previous research there can be no doubt as to the important effects of voice quality on the judgements that listeners will make about speakers. However the results of the further multi-dimensional scaling (table 3.8) are certainly interesting. The very strong correlations between the 2 dimensional representation of the eleven speakers in our study with their span and level measures would indicate that pitch range is certainly a potent characteristic that listeners tune into when making their judgements.

This project, though thorough, has only been small scale compared to the grand plan set out in Monaghan & Ladd (1990) as discussed on page 58, but its success has proven that a larger experiment is worthwhile. The next step then is to run a similar style experiment to that just described. The next experiment will investigate whether a similar style pitch range model can be found using English speech. And by making the experiment as large scale as possible, many further investigations can be made.

Not only will it be possible to see how much variation in pitch range can account for variation in listeners' judgements of speaker characteristics, the amount of variation due to other variables that have been unaccounted for in the Dutch experiment, such as the accent of speaker, the sex of speaker as well as the accent and sex of the listener can be accounted for. Also an investigation into which scale of measurement should be used for the pitch range model parameters - ERB, Hertz or musical semitones - will be carried out.

The Dutch experiment described has established one important fact; that the tonal targets used in the Shriberg *et al.* (1996) model to capture within-speaker variation, also can be used to capture differences in pitch range across speakers successfully. What is not yet established, and is to be a central part of this thesis, is whether the model captures these differences any more successfully than other suggested measures of pitch range which normally just include a "max-min" difference or some measure of the general distributional properties of f0 captured by the means and standard deviations for different speakers (as highlighted in section 2.1.1). It is this issue we turn to in the next chapter.

# Chapter 4

# Experiment 2

## 4.1 Introduction

In this chapter we report a large scale experiment designed to identify the best characterisation of pitch range for explaining paralinguistic effects. We go well beyond the pilot study based on Dutch (cf. chapter 3); we use nearly 3 times as many speakers and longer samples of speech. It will be seen that we reach very similar conclusions, namely that the best characterisation of pitch range for explaining listener judgements of speaker characteristics is one based on two partially independent variables, level and span, which are defined in terms of the scaling of linguistically defined targets in the pitch contours. Such a model consistently outperforms models based on long term distributional properties of f0 (LTD).

In section 2.1.1 we discussed various possible measures for pitch level and span. Two toplines and two bottomlines, relating to turning points in the f0 contour, were suggested, and we shall look at the measures of span and level in more detail in section 4.2.2. For the purposes of experiment 2, an average of a speaker's sentence-initial high (H) and an average of a speaker's non-sentence-initial highs (M) were taken as

the possible toplines; the average of a speaker's post-accent valleys (L) and the average of a speaker's sentence final low were used as measures of the potential bottomlines as well as potential measures of level. A more detailed discussion of the measurement points used to establish H, M, L and F appears in section 4.2.2.

The aim of experiment 2 is to establish which possible combination of H, M, L and F best represents span, and which measure, L or F, best represents the difference in level in accounting for the variation in listener judgements of speaker characteristics across speakers. Having established the most successful measure of span and level based on tonal targets in speech (which we are calling the linguistic measure), another key part of experiment 2 is the comparison of the linguistic measure with the other established measures of pitch range based on LTD properties of f0 for each speaker.

Another issue established in section 2.1.2 is the question of which scale to measure span and level. Experiment 2 examines whether differences in pitch range are best characterised using either the linear Hertz scale, the logarithmic musical semitone scale or the ERB-rate scale based on the frequency selectivity of the auditory system.

Because this chapter uses English rather than Dutch speakers, we are also able to investigate the interactions between pitch range and segmental characteristics (specifically regional accent) in their effects on listeners' judgements. Early work by Lambert (1972a) established the importance for listeners of the identification with members of their own linguistic group. Lambert (1972a) asked samples of French speaking and English speaking Montreal students to evaluate the personality characteristics of 10 speakers, some speaking in French, some speaking in English. The traits on which speakers were judged were *leadership, sense of humour, intelligence, religiousness, self-confidence, dependability, entertainingness, kindness, ambition, sociability, character* and *likability*. Bilinguals were used in the speech recordings: for 8 of the speech presentations only 4 different speakers were actually used. Subjects were in fact making judgements on the same set of speakers; the only difference was the language being spoken. Results showed that English subjects evaluated the English guises more favourably on

most traits. French subjects not only evaluated the English guises more favourably than French guises, but their evaluations of French guises were reliably less favourable than those of English subjects. Lambert (1972a) interprets this finding as evidence for a minority group reaction on the part of the French sample, and as a reflection of the influence of community-wide stereotypes of English and French speaking Canadians.

Cheyne (1970) looked at the evaluation of Scottish and English (to be regarded as representative of Southern Standard English - RP) voices and found that both Scottish and English listeners rated English male speakers (female speakers were not used) as possessing more leadership, intelligence, ambition and self confidence than Scottish speakers. The Scottish listeners showed some *accent loyalty*, evaluating their own group as more generous, goodhearted, friendly, humorous and likeable. The English listeners rated the Scottish voices as being more friendly. These results might be considered to be inconsistent with expectations. Generosity was one of the few scales where Scottish accents were judged more favourably, contradicting the more general stereotype of a Scottish trait for meanness. Cheyne (1970) suggests that "it is possible that stereotypes for speakers with regional accents are different from national stereotypes obtained by other means." Giles & Powesland (1975) sums up the findings of Cheyne (1970) concluding that "speakers of RP may attract stereotyped personality impressions of greater *competence* from listeners than speakers of nonstandard regional accents. This impression appears to transcend accent loyalty. However both regional accented judges and to a lesser extent RP judges seem to consider non standard speakers as possessing greater personality integrity and social attractiveness than RP speakers." The difference between the two different studies, Lambert (1972a) and Cheyne (1970), is that in the latter there is at least some accent loyalty.

The current study will investigate whether there is any accent loyalty in a similar fashion as the results of Cheyne (1970), interpreted for the current study's own selection of speaker characteristics, for two reasons. Firstly it is interesting to see if there are any major differences in how the accents of speakers and listeners interact in the judgement of speaker characteristics. Secondly, if there are major effects of regional accent

in the characterisation of speakers, this could be a confounding factor in the proposed methodology for evaluating measures of span and level. If for example an English person judges an English voice as being very confident and this can be attributed to the speaker's accent as opposed to a wide pitch span, the methodology selected for assessing span would not be as convincing.

Pitch range most conspicuously signals the difference between male and female voices. But, there are certain stereotypes held about men's and women's speech. Henton (1989) discusses stereotypes held about women's voices characterised as sounding like the "moo of a cow" and descriptions such as "high-pitched, shrill, over-emotional and swoopy." Henton (1989) suggests that labels for pitch are far more influenced by social evaluations of the speaker, than by the actual auditory level of the pitch itself. In her study of the effect of sex of speaker on pitch range used, Henton strictly controlled for confounding variables such as age, dialect, socio-economic status, etc. not normally well covered in similar studies. Henton's conclusions were that females do not employ a greater pitch range in English, and that female speech is ill-characterised as swoopy at least with regard to pitch range. A follow up study (Henton 1995) looking at pitch dynamism (defined as "the degree of rapidity of changes in a speaker's pitch range from high points to low points and vice versa" Henton 1995) concluded that this particular pitch feature also showed no significant difference in any of the conditions tested between females and males.

Chapter 4 also reports on secondary analyses which investigate the effects of sociolinguistic factors of speaker sex and accent and listener sex and accent on the characterisations of speakers' voices.

## 4.2 Stimulus Design and Analysis

### 4.2.1 Speakers

A total of 70 native speakers of English, 37 male and 33 female were recorded. About half of the speakers were students or employees at Edinburgh University. The remaining speakers were members of one of Edinburgh's amateur choral societies. None of these speakers were closely involved in research on prosody. The speakers were only informed of the nature of the study after each individual recording session had finished. For this experiment, as English speech was being used, it was necessary to control for effects of regional accent, as compared to the study described in chapter 3, which used Dutch speech. All speakers in the database had an accent mainly spoken by people from London and the Home Counties, or a Standard Scottish accent mainly spoken by people from the Lothian region. From this pool of 70 speakers, 32 speakers were used in the current study due to the limitations of time in the perception study to follow. The speech of 8 Scottish males, 8 Scottish females, 8 English males and 8 English females was used. In the current study there is again the assumption that there is enough variation in the voices to be able to make "ratable" differences in speaker characteristics, rather than relying on recordings of acted or simulated characteristics. The age of the speakers ranged between 19 and 67. A fairly even spread of ages were represented in each of the four groups of speakers. The ages of the speakers have been included in table 4.1.

### 4.2.2 Speech Materials

The basic approach used by Shriberg *et al.* (1996), to measure f0 at specific pre-selected points, was again utilised for the purposes of obtaining level and span measures in the current experiment. The f0 measures were used to establish stable mean pitch values for certain target levels creating the "map" of the relative pitch of these targets which could then be compared between speakers (as outlined in section 3.2.2).

The speech materials used in the recording session for Experiment 2 were 8 long passages that took about a minute each to read aloud. There was variation in the reading times due to speaking rate. The eight passages were selected from various linguistic text books, and from a selection of newspaper articles. Each passage was chosen on its simplicity to read and for being reasonably neutral for potential emotional effects on speakers' recording performance. Passages were chosen to elicit what can be described as normal speaking involvement from each subject; it was assumed that no one passage would make any speaker sound especially bored or especially excited.

Measurements were based on pitch data averaged from over whole passages. Measurements were taken at 4 selected target points in each sentence in each passage. These points are described as sentence initial high (H), non-initial accent peaks (M), post-accent valleys (L), and sentence final lows (F). For each sentence in a passage, by definition, there would be only one sentence initial high and one sentence final low, but there would be varying numbers of peaks and valleys depending on the length of each sentence. All tokens of each of the 4 types were collected into their respective category and then averaged to represent the data for each speaker for that particular target point. A description of the four pre-selected pitch targets are shown in figure 4.1.

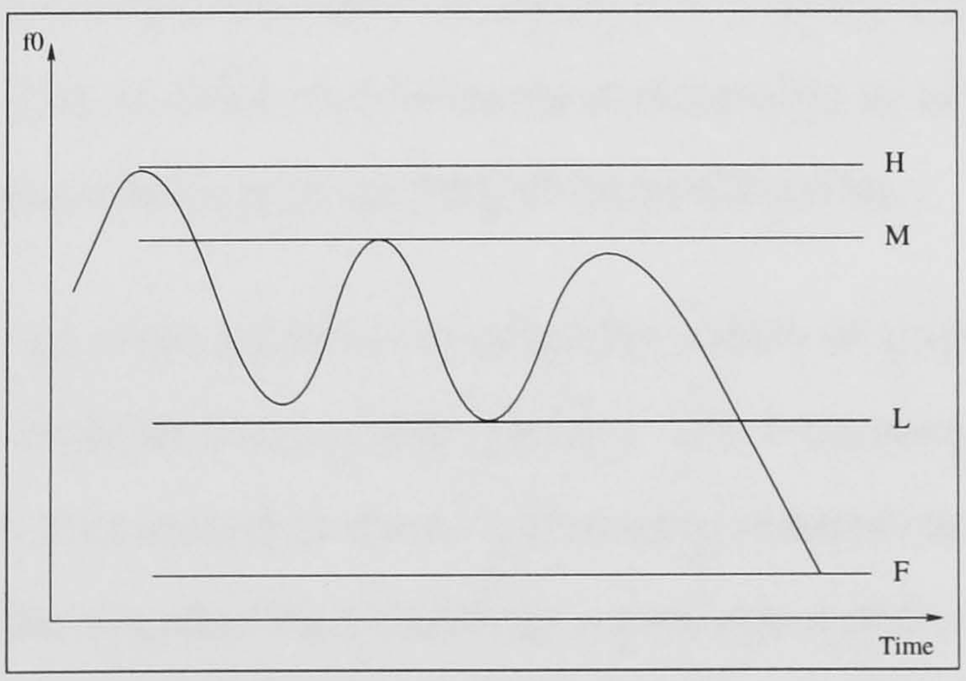



**Figure 4.1.** Measurement locations for span and level parameters on an idealised speaker contour

Horii (1975) found that for every script reading by a given talker, "statistical characteristics such as the standard deviations of the means converge fairly quickly, to within 2Hz of the total mean of a sample size of about 1 minute." As a minute of speech is considered a suitable amount of speech for investigations into the general distributional properties of f0 for a given speaker, this study assumes that over a minute of speech, measures of given linguistic targets such as sentence-initial peak will also be reasonably stable and valid.

### 4.2.3 Recordings

Each speaker recorded all the materials in a single recording session lasting about 15 minutes. Speech was recorded onto Digital Audio Tape (DAT). The data was transferred digitally from the DAT recorder to a Sun workstation and then copied to CD-ROM. The recordings were made in a quiet recording studio.

General instructions were given before the start of each recording session. Speakers were asked to make sure they were comfortable and then read the passages in a normal, natural style. No special instructions were given about the intonation to be used for the passages that have been used for this study. Before recording began, each subject was asked to read the first three sentences of the first passage. This gave an opportunity for the subjects to relax and be more comfortable in what, for many people, is a strange environment and a strange task to be involved in.

The passages to be read were printed on seperate sheets of paper, and placed all together, in order, on a table in front of the speaker. The recording technician, with the experimenter present, was seated in the neighbouring control room and controlled the presentation: when the speaker had finished a passage satisfactorily, the technician gave the speaker a short amount of time to look at the next passage, and then indicated when the recording was to start again. During the recording of each passage, if subjects felt that a sentence was not said correctly due to misreading or stumbling, they simply went to the beginning of the sentence in which the mistake had occurred

and read from there again. This happened only occasionally.

### 4.2.4 Pitch Range Analysis

Two of the 8 passages were chosen for use in the perception experiment, the MTV passage and the Railways passage, which can be found in appendix C, and only the speech recordings of those two passages were analysed for f0 data. The two passages that were selected were the fourth and the sixth to be recorded from the original eight passages. The two passages were selected on a number of criteria. The selected passages were the two that had the fewest speech errors, were the closest to a minute's worth of speech for each speaker, were far enough into the recording session so that the speakers were well settled into the task and not as far into the session so that the speakers would be showing any potential signs of boredom or being tired.

All the data for each measurement point was collected for each speaker using XWAVES, an interactive wave form processing package developed at Entropic Research Laboratory. Means for each of the measurement points were calculated. The results were initially kept separate for the two different passages just in case there were any notable differences in the readings by speakers. Results for the two passages by each speaker were similar so a single average for each point across the two passages was calculated. The two most stable measurement points were the sentence-initial high (H) and the sentence final low (F). For all speakers, the f0 scores for the two points, H and F, were based on 18 measurements, 9 from each passage. There was some variation in the number of measurements taken for the two other points, namely the non-initial accent peaks (M) and post-accent valleys (L). On occasion when no clear peak or valley was present, no measurement was taken. Essentially this means we were dealing with real peaks and real valleys and not textually defined measurement points chosen a priori. The measures for the M and L points for each speaker were based on roughly 70 data points. Across speakers there were fairly consistent patterns of accentuation, giving justification to the measurements made and the validity of our proposal that

we are measuring genuine target points in speech. In terms of variation in numbers of measurement points recorded, the average number of tokens of post accentual valleys across the 32 speakers was 69, with the range between 57 and 81.

Whilst the materials in the previous experiment (chapter 3) were designed to elicit specific target points, this was not the case in the current experiment. This is not to say that regular target points did not exist. The initial measurement procedure involved a study of 5 speakers to examine where there were clear and consistent peaks and valleys across speakers. The criteria for specific measurement points were the same as described in section 3.2.4. It is assumed that the H and F target points are easy to find on an f0 contour. Clearly there was some variation in the number of M and L target points as pointed out above. Only clear and major turning points at the top and bottom of a speaker's f0 contour were selected for the M and L target points. Minor turning points were considered to be due to segmental perturbations. For a more detailed look at how decisions were made on measurement locations, waveforms and f0 contours for 3 speakers (JB, JW and NC) uttering the phrase "*The project is the latest brainchild of the Planet Hollywood stable*" can be found in Appendix B, along with a commentary on the measurement process.

The details of the long term distribution of f0 were extracted from the two passages that were to be used as part of the perception experiment. The LTD data was compiled automatically using XWAVES. All extracted f0 values were used, not just selected peaks and valleys. The mean, standard deviation, kurtosis, skew, maximum and minimum values, and various percentile values were extracted using the SPSS statistics package. These LTD details were necessary to test the success at capturing the differences of listener judgements by various other suggested ways of measuring pitch range based on the LTD of f0.

| | | | | Span (ERB) | | | | Level (ERB) | |
|---|---|---|---|---|---|---|---|---|---|
| Speaker | Age | Sex | Accent | H-F | H-L | M-F | M-L | L | F |
| AP | 22 | male | Scottish | 2.85 | 2.42 | 1.80 | 1.38 | 3.55 | 3.13 |
| HM | 24 | male | Scottish | 2.28 | 1.81 | 1.45 | 0.98 | 3.79 | 3.32 |
| JS | 33 | male | Scottish | 2.67 | 2.37 | 1.80 | 1.49 | 3.26 | 2.95 |
| AB | 35 | male | Scottish | 2.51 | 2.07 | 1.77 | 1.32 | 3.55 | 3.10 |
| GF1 | 43 | male | Scottish | 2.52 | 2.16 | 1.55 | 1.19 | 3.37 | 3.01 |
| FH | 50 | male | Scottish | 2.33 | 2.01 | 1.44 | 1.12 | 3.30 | 2.98 |
| TM | 54 | male | Scottish | 1.94 | 1.58 | 1.32 | 0.96 | 3.07 | 2.71 |
| KW | 67 | male | Scottish | 2.53 | 1.79 | 2.09 | 1.35 | 3.13 | 2.39 |
| SM | 19 | male | English | 2.49 | 2.05 | 1.38 | 0.94 | 3.41 | 2.97 |
| GF2 | 21 | male | English | 1.05 | 0.94 | 0.67 | 0.55 | 3.46 | 3.35 |
| RC | 25 | male | English | 2.34 | 2.10 | 1.24 | 1.00 | 3.25 | 3.00 |
| ME | 33 | male | English | 2.55 | 2.28 | 1.47 | 1.20 | 3.28 | 3.01 |
| RL | 37 | male | English | 2.00 | 1.71 | 1.27 | 0.98 | 3.31 | 3.02 |
| VR | 45 | male | English | 2.72 | 2.11 | 2.09 | 1.48 | 2.87 | 2.27 |
| JB | 54 | male | English | 3.09 | 2.33 | 2.48 | 1.71 | 3.63 | 2.86 |
| GB | 65 | male | English | 3.01 | 2.29 | 2.24 | 1.51 | 3.38 | 2.66 |
| FL | 21 | female | English | 1.61 | 1.18 | 1.04 | 0.61 | 5.05 | 4.63 |
| NG | 22 | female | English | 2.05 | 1.78 | 1.14 | 0.81 | 5.47 | 5.20 |
| SO | 24 | female | English | 3.80 | 3.26 | 1.93 | 1.40 | 5.34 | 4.80 |
| NC | 35 | female | English | 3.78 | 3.21 | 2.38 | 1.81 | 4.69 | 4.12 |
| JK | 41 | female | English | 2.90 | 2.14 | 2.05 | 1.30 | 5.27 | 4.51 |
| JV | 47 | female | English | 1.79 | 1.24 | 1.28 | 0.72 | 5.21 | 4.65 |
| RS | 54 | female | English | 2.64 | 2.25 | 1.72 | 1.33 | 4.45 | 4.06 |
| MT | 60 | female | English | 3.91 | 3.41 | 2.32 | 1.81 | 4.36 | 3.85 |
| JT | 20 | female | Scottish | 1.84 | 1.62 | 1.16 | 0.95 | 5.78 | 5.56 |
| JC | 21 | female | Scottish | 1.67 | 1.08 | 1.18 | 0.59 | 5.58 | 4.99 |
| KG | 21 | female | Scottish | 2.69 | 2.43 | 1.80 | 1.54 | 4.96 | 4.70 |
| DN | 39 | female | Scottish | 1.86 | 1.36 | 1.41 | 0.91 | 5.18 | 4.68 |
| SS | 36 | female | Scottish | 3.94 | 2.97 | 2.64 | 1.67 | 5.21 | 4.24 |
| AW | 53 | female | Scottish | 3.73 | 3.25 | 2.52 | 2.04 | 5.17 | 4.69 |
| JD | 62 | female | Scottish | 2.62 | 2.28 | 1.48 | 1.13 | 5.04 | 4.69 |
| JO | 66 | female | Scottish | 2.95 | 2.60 | 1.95 | 1.61 | 4.67 | 4.33 |

**Table 4.1.** Span and level measures for each speaker from the English Speech Database, measured in ERB

| | | | Span (ERB) | | | Level (ERB) | |
|---|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | meanf0±2sds | 95 - 5%f0 | 90 - 10%f0 | meanf0 | medianf0 |
| AP | male | Scottish | 4.34 | 3.30 | 2.60 | 4.30 | 4.20 |
| HM | male | Scottish | 3.90 | 3.72 | 2.20 | 4.20 | 4.14 |
| JS | male | Scottish | 3.90 | 3.18 | 2.40 | 3.87 | 3.70 |
| AB | male | Scottish | 3.52 | 2.95 | 2.38 | 4.28 | 4.25 |
| GF1 | male | Scottish | 4.29 | 3.79 | 3.15 | 3.86 | 3.91 |
| FH | male | Scottish | 3.36 | 2.69 | 2.05 | 3.81 | 3.62 |
| TM | male | Scottish | 2.97 | 2.46 | 1.82 | 3.44 | 3.34 |
| KW | male | Scottish | 3.65 | 3.07 | 2.42 | 3.69 | 3.59 |
| SM | male | English | 3.60 | 3.13 | 2.32 | 3.95 | 3.88 |
| GF2 | male | English | 2.96 | 2.38 | 1.38 | 3.78 | 3.70 |
| RC | male | English | 3.41 | 2.62 | 2.06 | 3.76 | 3.61 |
| ME | male | English | 3.73 | 3.18 | 2.16 | 3.90 | 3.80 |
| RL | male | English | 3.37 | 2.66 | 1.99 | 3.79 | 3.69 |
| VR | male | English | 3.68 | 3.05 | 2.50 | 3.44 | 3.28 |
| JB | male | English | 4.79 | 4.26 | 3.59 | 4.25 | 4.20 |
| GB | male | English | 3.90 | 3.44 | 2.71 | 4.08 | 3.99 |
| FL | female | English | 4.57 | 4.38 | 3.29 | 5.23 | 5.32 |
| NG | female | English | 4.64 | 4.50 | 2.39 | 5.80 | 5.76 |
| SO | female | English | 5.76 | 4.90 | 3.29 | 5.91 | 5.78 |
| NC | female | English | 5.96 | 5.85 | 3.92 | 5.51 | 5.44 |
| JK | female | English | 5.08 | 4.77 | 3.00 | 5.81 | 5.83 |
| JV | female | English | 4.85 | 4.53 | 3.91 | 5.34 | 5.46 |
| RS | female | English | 4.82 | 4.10 | 3.00 | 5.08 | 5.03 |
| MT | female | English | 5.66 | 4.72 | 3.78 | 5.21 | 4.96 |
| JT | female | Scottish | 4.65 | 4.33 | 2.33 | 6.10 | 6.12 |
| JC | female | Scottish | 4.85 | 4.55 | 3.69 | 5.66 | 5.79 |
| KG | female | Scottish | 5.02 | 4.02 | 3.16 | 5.64 | 5.53 |
| DN | female | Scottish | 4.15 | 4.16 | 2.05 | 5.61 | 5.68 |
| SS | female | Scottish | 6.04 | 5.85 | 3.80 | 5.74 | 5.70 |
| AW | female | Scottish | 5.95 | 4.82 | 3.80 | 6.02 | 5.88 |
| JD | female | Scottish | 5.00 | 4.87 | 2.79 | 5.42 | 5.40 |
| JO | female | Scottish | 5.09 | 4.11 | 3.36 | 5.38 | 5.30 |

**Table 4.2.** Long term distributional properties of f0 for each speaker from the English Speech Database, measured in ERB

## 4.3 Pitch Range Results

Figure 4.1 shows a schematic drawing of an idealised f0 contour with a visual description of the measurement points that were used for modelling speakers' pitch range. It should be clear form this figure that H and M indicate potential toplines of the speaker span, L and F indicate potential bottomlines of the speaker span as well as potential measures for level. Only the bottomline was considered a measure of level in keeping with Shriberg *et al.* (1996). The bottomline is a good stable measurement point to capture cross-speaker differences in level which is not affected greatly by variations in span.

The set of data shown in table 4.1 are the averaged pitch range results for each speaker. In keeping with the presentation in section 3.3, these results are all represented using the ERB-rate scale. The full set of results, which include the results for the Hertz, ERB-rate and semitone scales, are in appendix A from table A.1 to table A.3.

The set of data shown in table 4.2 are measures of span and level taken from the long term distributional properties of f0 for each speaker. The measures represented in table 4.2 have all been used in previous research to represent span and level. For level, Kraayeveld (1997) uses mean f0 and Bezooijen (1984) uses median f0. For span, Jassem (1971) uses four standard deviations around the mean, Horii (1975) uses the difference between the 95th and the 5th percentile (which accounts for a 90% range of f0) and Williams & Stevens (1972) uses the difference betweeen the 90th and 10th percentile (which accounts for an 80% range of f0). The results for the level and span of speakers based on LTD properties are all represented using the ERB-rate scale. The full set of results for the Hertz, ERB-rate and semitone scales are in appendix A from table A.4 to table A.6.

From table 4.1 it can be seen that VR has the lowest level using either F or L as the measure. JT has the highest level using either the F or L measure. For span, across all the 4 measurment options, GF2 has the narrowest span. SS has the widest span for the

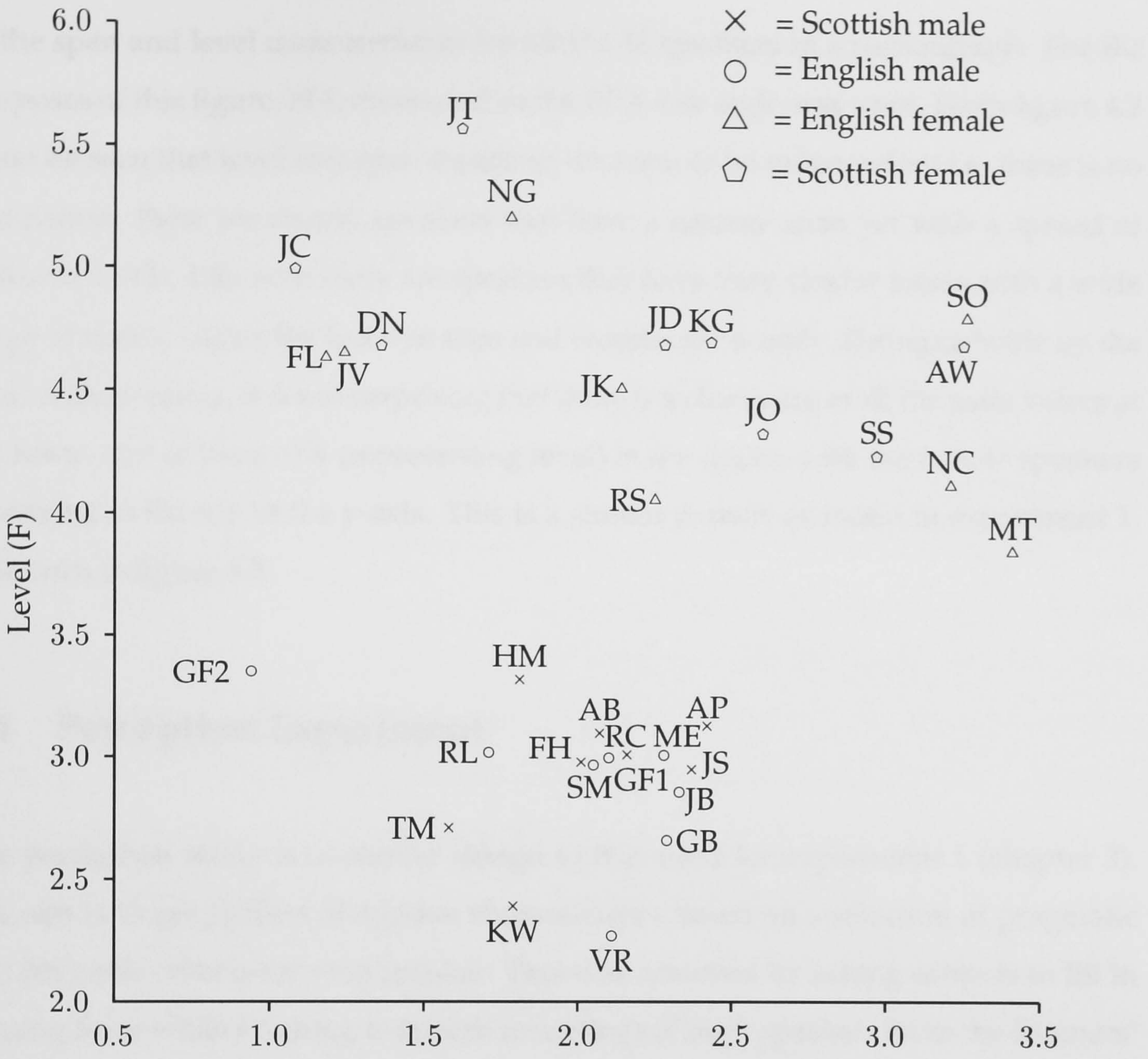



**Figure 4.2.** Variations in span and level of the 32 speakers of English

H-F and the M-F measure of span, MT has the widest span for the H-L measure and AW has the widest span for the M-L measure.

From table 4.2 it can be seen for level that, again, VR has the lowest level and JT has the highest level for both measures. GF2 has the narrowest span for all the measures. SS has the widest span for the four standard deviations around the mean measure, SS and NC have the widest span using the 90% range and NC has the widest span using the 80% range.

Figure 4.2 shows some of the information from table 4.1, giving a visual representation

of the span and level measurements for all the 32 speakers in a scattergraph. For the purposes of this figure, H-L measured on the ERB-rate scale was used. From figure 4.2 it can be seen that level and span measures do seem to be independent i.e. there is no correlation: there are clearly speakers that have a narrow span yet with a spread of different levels. Likewise there are speakers that have very similar levels with a wide range of spans. Given the fact that men and women are readily distinguishable by the level of their voices, it is not surprising that there is a clustering of all the male voices at the lower end of the y-axis (representing level) in the figure, with the female speakers clustering at the top of the y-axis. This is a similar pattern as found in experiment 1, as shown in figure 3.5.

## 4.4 Perception Experiment

The perception study is of similar design to that used for experiment 1 (chapter 3). The aim is to get profiles of speaker characteristics based on a selection of pragmatic and phonetic criteria for each speaker. This was achieved by asking subjects to fill in a rating form while listening to speech recordings of each speaker. From the listeners' responses it is possible to see how each speaker can be characterised. With this data it is possible to compare the perception of speaker characteristics with the data showing across-speaker variation in pitch range. This was done in chapter 3 to show that linguistic measures of pitch range show similar patterns found in research that have used long term distributional properties of f0 in speaker characteristics research (cf. Brown *et al.* 1973, Bezooijen 1984). For the current experiment all the various suggestions for measures of span and level will be correlated with the speaker characteristics data to assess which measure best characterises the variation of level and span in relation to speaker characterisitics.

There is one further development in the current study compared to the small scale

study discussed in chapter 3. The speech recordings were not only presented to speakers in their normal condition, the MTV and Railways speech recordings were also passed through a low pass filter.

Low-pass filtering is a technique which filters out the verbal content and voice quality, but leaves the fundamental frequency unaltered. Scherer *et al.* (1984) states that "commonly used electronic content-filtering techniques use a single cutoff frequency of about 500 Hz, with a rolloff of between 30 and 40 dB/oct. While this destroys intelligibility, it is probable that it still leaves some voice quality information in the signal." Therefore Scherer *et al.* (1984) set the cutoff frequency for each utterance at each utterance's own highest f0 value. For the speech in the current experiment various options for setting the level of the cutoff were tried. Using the 500 Hz cutoff , verbal content of the speech was still reasonably clear. Using the highest f0 value for each speaker on the otherhand made the speech so muffled that the task of making speaker characteristic judgements would be too difficult and frustrating. For the current experiment a compromise cutoff point was used. The highest f0 value for each speaker was measured and this level was raised by 4 semitones (a musical major third) with a ceiling of 500 Hz, always with a 60 dB/oct rolloff. This ceiling affected only speaker SO. The highest f0 for each speaker, and the resulting cutoff points for low-pass filtering can be found in appendix D.

The whole experiment is based on the latin square design. For this experiment there were four different groups. The 32 speakers appear in each group only once. For example speaker 1 reading the MTV passage was in group 1, speaker 1 reading the Railways passage was in group 2, speaker 1 reading the MTV passage which had been low-pass filtered was in group three and speaker 1 reading the Railways passage which had been low-pass filtered was in group four. To complete the 4 X 4 design required for a latin square, the other cells were divided into the four sex/accent differences between the speakers as shown in table 4.3. For the current experiment, there were 8 speakers in each cell of the latin square.

| | MTV Norm | MTV LPF | Railways Norm | Railways LPF |
|---|---|---|---|---|
| Male Scot. | 8 | 8 | 8 | 8 |
| Male Eng. | 8 | 8 | 8 | 8 |
| Female Eng. | 8 | 8 | 8 | 8 |
| Female Scot. | 8 | 8 | 8 | 8 |

**Table 4.3.** Latin square design of the perception experiment, with 8 speakers in each cell

### 4.4.1 Speech Materials

The full compliment of 32 speakers was used, 8 Scottish females (JT, JC, KG, DN, SS, AW, JD, JO), 8 English females (FL, NG, SO, NC, JK, JV, RS, MT), 8 Scottish males (AP, HM, JS, AB, GF1, FH, TM, KW) and 8 English males (SM, GF2, RC, ME, RL, VR, JB, GB). The speech materials used were the two passages that were analysed in the acoustic study.

Four experimental tapes were made up. On each tape was all the speech material for a full experimental run. This consisted of one of four possible speech presentations for each speaker as well as a presentation of two speakers that were used for a practice run. The trial run was used so the listener judges had the opportunity to become familiar with the task that was being asked of them. The stimuli were played to groups of subjects on a high quality tape machine with high quality speakers in a large room.

Each presentation of a speaker consisted of one passage. The two different passages were two of the possible presentations. The other two presentations were based on the output of the two selected passages having been passed through the low-pass filter.

The speakers were put in pseudo-random order. Firstly it was always ensured that each of the four types of passage presentation occurred once in every four stimuli. In each of the four different experimental conditions, the speakers were presented in different orders.

### 4.4.2 Listener Judges

There were 48 subjects in 9 experimental sessions. All of the subjects were linguistic students taken from first or second year undergraduate level. For the four different experimental conditions, 12 subjects were in each group of which 6 were Scottish and 6 were English.

### 4.4.3 Rating Forms

For the current experiment there were 12 features to be judged - *confident, tense, harsh, expressive, deep, weak, irritated, happy, afraid, relaxed, emphatic* and *bored*. The same 7 point unipolar scale used in experiment 1 was used for this experiment (cf. section 3.4.1). The reasons for dropping 8 of the features from the previous experiment were partly due to time constraint on the experimental session, and partly based on results of the previous experiment. The focus of the experiment in this chapter is centred on our pitch range study, moving away from basing our research on previous studies (e.g. Uldall 1964, Brown *et al.* 1973). The main bulk of those features dropped were the voice quality features (*whisper, breathy, creaky, nasal*) though two (*tense* and *harsh*) were kept in the current study to act as control items, as in the previous experiment. *Sad* and *unemphatic* were dropped because they were only acting as opposites for *happy* and *emphatic*, providing a control to establish the validity of our methodology. *Deep* was again maintained as a good control feature for level measures.

### 4.4.4 Experimental Session

Subjects were given a rating form booklet. On the cover of the booklet were general instructions as to how the experiment would be run, as well as two examples of the rating form. Once they had finished reading the instructions, the subjects had an opportunity to ask questions. Then a practice run was carried out using the speech

of two different speakers, one being the normal presentation of the MTV passage, the other being an example of the Railways passage having been passed through a low pass filter. After a further opportunity for questions, the experiment proper was run. The whole experimental session lasted for one hour.

## 4.5 Results

### 4.5.1 Primary analyses

The *mode* for each feature for each speaker averaging across all listeners was calculated using the SPSS statistical package. The full results, including the mode for each feature for each speaker averaging across all listeners for each of the four seperate conditions (passage x 2, filtering condition x 2), can be found in appendix E. Modes for each feature for each speaker for the MTV passage in the normal speech condition are also shown in table 4.4. Modes for each feature for each speaker for the MTV passage in the low-pass filtered condition are also shown in table 4.5. For example, tables 4.4 and 4.5 show that speaker AP, a Scottish male, was judged as being "5" on the confident scale in both the normal and the low-pass filtered speech condition for the MTV passage. So according to the mode score, generally the listeners perceived speaker AP as sounding reasonably confident. Speaker MT, an English female, was judged as being "1" on the bored scale for the normal speech and "2" on the bored scale for the low-pass filtered speech for the MTV passage. So according to the mode score, generally the listeners perceived speaker MT as sounding not at all bored when the full verbal content in the speech could be heard, yet when the speech was degraded the listeners generally perceived speaker MT to be sounding slightly bored.

For the next stage the relationships between the pitch range parameters for each speaker (table A.1 through to table A.6) and the results of the perception study (table E.1 through to table E.4) were established by calculating Spearman's rank correlation coefficients ($\rho$). Table 4.6 shows the full correlation coefficient results comparing the

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 5 | 4 | 4 | 6 | 3 | 4 | 5 | 5 |
| tense | 2 | 3 | 2 | 1 | 6 | 1 | 2 | 4 |
| harsh | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 |
| expressive | 3 | 2 | 3 | 3 | 4 | 3 | 5 | 3 |
| deep | 4 | 3 | 5 | 3 | 3 | 3 | 6 | 5 |
| weak | 2 | 5 | 3 | 2 | 2 | 1 | 2 | 1 |
| irritated | 3 | 2 | 1 | 2 | 2 | 2 | 1 | 2 |
| happy | 1 | 2 | 5 | 2 | 2 | 3 | 2 | 2 |
| afraid | 2 | 1 | 3 | 4 | 1 | 1 | 2 | 1 |
| relaxed | 2 | 2 | 2 | 1 | 1 | 6 | 5 | 3 |
| emphatic | 3 | 2 | 5 | 3 | 6 | 2 | 5 | 3 |
| bored | 5 | 3 | 2 | 1 | 2 | 3 | 2 | 3 |
| | English Men | | | | | | | |
| aka | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 6 | 5 | 3 | 7 | 3 | 4 | 6 | 5 |
| tense | 1 | 5 | 5 | 1 | 5 | 1 | 2 | 2 |
| harsh | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| expressive | 5 | 2 | 3 | 5 | 5 | 5 | 6 | 3 |
| deep | 4 | 5 | 3 | 5 | 3 | 5 | 3 | 4 |
| weak | 2 | 3 | 5 | 1 | 6 | 1 | 2 | 2 |
| irritated | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| happy | 5 | 2 | 2 | 3 | 3 | 4 | 2 | 2 |
| afraid | 1 | 1 | 5 | 1 | 3 | 1 | 1 | 1 |
| relaxed | 5 | 3 | 2 | 5 | 2 | 5 | 5 | 5 |
| emphatic | 5 | 3 | 3 | 2 | 4 | 2 | 5 | 5 |
| bored | 3 | 3 | 3 | 2 | 2 | 1 | 2 | 1 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 3 | 4 | 3 | 5 | 5 | 3 | 5 | 7 |
| tense | 5 | 2 | 4 | 1 | 5 | 4 | 5 | 3 |
| harsh | 4 | 1 | 5 | 3 | 2 | 2 | 2 | 4 |
| expressive | 1 | 3 | 4 | 5 | 2 | 2 | 4 | 3 |
| deep | 3 | 1 | 3 | 2 | 2 | 2 | 3 | 3 |
| weak | 3 | 3 | 2 | 1 | 3 | 4 | 2 | 2 |
| irritated | 4 | 1 | 1 | 1 | 4 | 5 | 3 | 2 |
| happy | 1 | 5 | 4 | 5 | 1 | 1 | 2 | 3 |
| afraid | 3 | 1 | 2 | 1 | 1 | 1 | 3 | 1 |
| relaxed | 1 | 5 | 2 | 5 | 3 | 2 | 2 | 2 |
| emphatic | 1 | 3 | 5 | 4 | 4 | 1 | 3 | 6 |
| bored | 6 | 2 | 2 | 2 | 3 | 2 | 2 | 1 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 4 | 2 | 4 | 3 | 5 | 7 | 5 | 5 |
| tense | 2 | 2 | 4 | 5 | 2 | 2 | 5 | 3 |
| harsh | 1 | 1 | 2 | 2 | 1 | 4 | 2 | 1 |
| expressive | 4 | 2 | 5 | 3 | 6 | 4 | 4 | 2 |
| deep | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 3 |
| weak | 3 | 2 | 2 | 5 | 2 | 1 | 3 | 1 |
| irritated | 2 | 4 | 1 | 2 | 1 | 1 | 3 | 1 |
| happy | 5 | 2 | 5 | 2 | 5 | 3 | 2 | 2 |
| afraid | 5 | 1 | 3 | 2 | 2 | 1 | 5 | 2 |
| relaxed | 2 | 1 | 2 | 1 | 4 | 3 | 2 | 2 |
| emphatic | 4 | 2 | 5 | 2 | 5 | 5 | 3 | 4 |
| bored | 2 | 7 | 3 | 3 | 3 | 3 | 4 | 2 |

**Table 4.4.** Mode results of normal speech for all speakers reading the MTV passage

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 5 | 3 | 5 | 3 | 2 | 5 | 5 | 5 |
| tense | 2 | 5 | 3 | 5 | 2 | 2 | 2 | 3 |
| harsh | 1 | 1 | 5 | 2 | 2 | 1 | 4 | 2 |
| expressive | 3 | 3 | 5 | 4 | 3 | 4 | 2 | 4 |
| deep | 4 | 5 | 5 | 6 | 5 | 4 | 6 | 7 |
| weak | 1 | 2 | 1 | 3 | 3 | 2 | 1 | 1 |
| irritated | 2 | 1 | 2 | 3 | 3 | 1 | 3 | 2 |
| happy | 1 | 2 | 5 | 2 | 1 | 2 | 1 | 2 |
| afraid | 2 | 5 | 1 | 1 | 2 | 2 | 1 | 2 |
| relaxed | 5 | 3 | 6 | 2 | 3 | 5 | 2 | 4 |
| emphatic | 3 | 3 | 5 | 4 | 3 | 4 | 3 | 5 |
| bored | 2 | 6 | 1 | 3 | 1 | 4 | 4 | 2 |
| | English Men | | | | | | | |
| | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 3 | 2 | 4 | 5 | 6 | 5 | 5 | 5 |
| tense | 2 | 2 | 2 | 2 | 4 | 3 | 5 | 2 |
| harsh | 3 | 1 | 2 | 2 | 2 | 1 | 3 | 2 |
| expressive | 5 | 2 | 2 | 5 | 2 | 3 | 5 | 5 |
| deep | 5 | 5 | 4 | 6 | 4 | 2 | 2 | 5 |
| weak | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 3 |
| irritated | 2 | 5 | 3 | 3 | 3 | 1 | 3 | 2 |
| happy | 3 | 1 | 2 | 2 | 4 | 3 | 3 | 3 |
| afraid | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 |
| relaxed | 5 | 5 | 5 | 4 | 4 | 5 | 3 | 6 |
| emphatic | 5 | 3 | 2 | 5 | 2 | 4 | 3 | 6 |
| bored | 1 | 5 | 3 | 3 | 5 | 3 | 3 | 3 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 2 | 3 | 5 | 5 | 2 | 2 | 3 | 5 |
| tense | 3 | 5 | 5 | 2 | 5 | 5 | 3 | 2 |
| harsh | 2 | 1 | 5 | 2 | 1 | 2 | 1 | 2 |
| expressive | 2 | 3 | 4 | 5 | 3 | 3 | 4 | 5 |
| deep | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
| weak | 4 | 5 | 3 | 2 | 3 | 5 | 4 | 2 |
| irritated | 4 | 3 | 3 | 2 | 3 | 2 | 6 | 2 |
| happy | 1 | 1 | 2 | 4 | 2 | 1 | 2 | 5 |
| afraid | 2 | 4 | 3 | 2 | 3 | 4 | 1 | 1 |
| relaxed | 3 | 2 | 2 | 5 | 2 | 3 | 3 | 5 |
| emphatic | 2 | 2 | 4 | 6 | 3 | 4 | 2 | 5 |
| bored | 3 | 3 | 2 | 2 | 2 | 5 | 5 | 2 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 3 | 2 | 5 | 5 | 5 | 5 | 5 | 5 |
| tense | 5 | 5 | 2 | 3 | 2 | 3 | 5 | 2 |
| harsh | 2 | 1 | 1 | 3 | 1 | 2 | 5 | 3 |
| expressive | 2 | 2 | 4 | 2 | 5 | 4 | 2 | 5 |
| deep | 1 | 1 | 2 | 2 | 1 | 3 | 4 | 3 |
| weak | 6 | 1 | 1 | 2 | 2 | 2 | 3 | 2 |
| irritated | 2 | 2 | 3 | 2 | 1 | 3 | 4 | 2 |
| happy | 2 | 1 | 3 | 2 | 3 | 4 | 2 | 2 |
| afraid | 3 | 5 | 1 | 3 | 1 | 2 | 1 | 2 |
| relaxed | 2 | 1 | 3 | 2 | 5 | 3 | 1 | 5 |
| emphatic | 4 | 2 | 2 | 3 | 5 | 4 | 2 | 4 |
| bored | 1 | 3 | 5 | 3 | 2 | 3 | 3 | 5 |

**Table 4.5.** Mode results of low pass filtered speech for all speakers reading the MTV passage

normal and low-pass filtered mode results with the linguistic measures of span and level measured in ERB. All the correlation coefficients that are significant ($p < 0.05$) are in a bold font. For example, table 4.6 show that the features *deep* and *relaxed* correlate with the level parameter L in the normal speech condition, and the features *confident, tense, deep, weak, afraid* and *relaxed* correlate with the level parameter L in the filtered condition.

L and F are considered as competing linguistic measures for level, and M-L, M-F, H-L and H-F are considered as competing linguistic measures for span. As can be seen in table 4.6 by the coefficients marked with a tick to indicate the measure showing the strongest correlation with each respective feature, the span measure M-L shows the strongest correlations for the most features for both the normal and filtered speech. Results for level are not so conclusive. Both L and F show strong correlations with the feature *deep*, as would be essential for any effective measure of level.

It is clear that for some speaker characteristics, notably *confident* and *bored*, effects of level and span are partially independent. This supports the hypothesis that two linguistically motivated, partially independent dimensions of variation better characterise the communicative effects of pitch range compared to the single dimension of just max-min f0. As correlation results are similar for both the normal and the filtered speech, these results lend support to the claim that there is a genuine independent pitch range effect in the characterisation of speakers.

The full set of results of the correlation analyses can be found in appendix F. The results in appendix F are set out in a similar fashion to table 4.6. The tables in appendix F show results for all possible measures of level and span including suggested long term distributional measures of span and level, for both the MTV and Railways passages. There are also results for the different scales for measuring level and span, namely Hertz, ERB and semitones.

A summary of the results in appendix F shows that for the competing measures of

| Feature | Normal Speech Level L | | Level F | | Span M-L | | Span M-F | | Span H-L | | Span H-F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| confident | -0.187 | | -0.279 | | **0.512** | | **0.528** | √ | **0.464** | | **0.507** | |
| tense | 0.169 | | 0.274 | | -0.272 | | -0.257 | | -0.203 | | -0.227 | |
| harsh | 0.184 | | 0.199 | | 0.269 | | 0.270 | | **0.305** | √ | 0.285 | |
| expressive | -0.212 | | -0.299 | | **0.411** | √ | **0.350** | | **0.396** | | **0.402** | |
| deep | **−0.834** | √ | **−0.802** | | 0.024 | | 0.001 | | -0.041 | | -0.027 | |
| weak | 0.241 | | 0.275 | | **−0.583** | √ | **−0.580** | | **−0.485** | | **−0.523** | |
| irritated | 0.294 | | 0.246 | | **−0.516** | √ | **−0.353** | | **−0.487** | | **−0.430** | |
| happy | 0.029 | | 0.077 | | 0.292 | | 0.154 | | **0.343** | √ | 0.278 | |
| afraid | 0.074 | | -0.230 | | -0.063 | | -0.235 | | 0.003 | | -0.122 | |
| relaxed | **−0.309** | | **−0.427** | √ | 0.245 | | 0.278 | | 0.169 | | 0.287 | |
| emphatic | 0.045 | | -0.114 | | **0.564** | | **0.496** | | **0.603** | √ | **0.585** | |
| bored | 0.255 | | **0.305** | √ | **−0.346** | √ | -0.273 | | -0.258 | | -0.288 | |
| | Filtered Speech Level L | | Level F | | Span M-L | | Span M-F | | Span H-L | | Span H-F | |
| confident | **−0.328** | √ | -0.287 | | **0.598** | √ | **0.517** | | **0.529** | | **0.525** | |
| tense | **0.487** | √ | **0.408** | | -0.264 | | -0.166 | | -0.295 | | -0.223 | |
| harsh | -0.125 | | -0.117 | | 0.103 | | 0.077 | | 0.163 | | 0.100 | |
| expressive | -0.138 | | **−0.297** | √ | **0.734** | √ | **0.713** | | **0.718** | | **0.731** | |
| deep | **−0.772** | √ | **−0.681** | | -0.085 | | -0.078 | | -0.146 | | -0.189 | |
| weak | **0.426** | √ | **0.386** | | -0.185 | | -0.190 | | -0.098 | | -0.098 | |
| irritated | 0.070 | | 0.211 | | -0.188 | | -0.282 | | -0.086 | | -0.161 | |
| happy | -0.168 | | -0.235 | | **0.692** | √ | **0.620** | | **0.604** | | **0.625** | |
| afraid | **0.542** | √ | **0.436** | | -0.282 | | -0.179 | | -0.259 | | -0.221 | |
| relaxed | **−0.553** | | **−0.589** | √ | **0.392** | √ | **0.318** | | **0.341** | | **0.337** | |
| emphatic | -0.209 | | **−0.360** | √ | **0.480** | | **0.548** | √ | **0.410** | | **0.469** | |
| bored | -0.076 | | 0.049 | | -0.235 | | **−0.301** | | **−0.310** | | **−0.341** | √ |

**Table 4.6.** Results of correlation analyses for 2 linguistic measures of level and 4 linguistic measures of span (measured in ERB) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

span, the M-L measure (the difference between average non-initial peak and average post-accent valley) generally shows the strongest correlations with listener judgements of speaker characteristics. Results for the competing measures of level show that F (average sentence final low) shows the strongest correlations with listener judgements of speaker characteristics.

The results for L and F are the same, whether level is measured using Hz or ERB, so for further study F measured in Hz will be used as the measure of level.

| | Span M-L | | | | | |
|---|---|---|---|---|---|---|
| | Normal Speech | | | Filtered Speech | | |
| Feature | Hertz | ERB | Semitone | Hertz | ERB | Semitone |
| confident | **0.464** | **0.512** | **0.529** √ | **0.486** | **0.598** | **0.609** √ |
| tense | -0.158 | -0.272 | **−0.410** √ | -0.113 | -0.264 | **−0359** √ |
| harsh | **0.307** √ | 0.269 | 0.114 | 0.076 | 0.103 | 0.118 |
| expressive | **0.355** | **0.411** | **0.413** √ | **0.647** | **0.734** √ | **0.687** |
| deep | -0.228 | 0.024 | **0.333** √ | **−0.311** √ | -0.085 | 0.196 |
| weak | **−0.470** | **−0.583** | **−0.632** √ | -0.007 | -0.185 | **−0.309** √ |
| irritated | **−0.400** | **−0.516** | **−0.546** √ | -0.109 | -0.188 | -0.258 |
| happy | **0.314** √ | 0.292 | 0.237 | **0.614** | **0.692** √ | **0.640** |
| afraid | 0.022 | -0.063 | -0.160 | -0.150 | -0.282 | **−0.389** √ |
| relaxed | 0.143 | 0.245 | **0.341** √ | 0.178 | **0.392** | **0.545** √ |
| emphatic | **0.558** | **0.564** √ | **0.465** | **0.384** | **0.480** | **0.543** √ |
| bored | -0.276 | **−0.346** | **−0.453** √ | -0.288 | -0.235 | -0.249 |

**Table 4.7.** Comparison of the correlation coefficients between the M-L span, using the 3 different measurement scales, and listener judges' ratings of the MTV passage for 12 adjectives for both normal and filtered speech.

Table 4.7 shows a comparison of the different ways of measuring the M-L span for the normal and low-pass filtered speech of the MTV passage. It is clear that all three types of measure (Hertz, ERB, semitones) capture differences between speakers' pitch span effectively, but generally speaking the ERB and the semitone measures show stronger correlations with listener judges than the Hz scale. There is not much difference between the ERB and the semitone measures, though the semitone measure is slightly more successful for both the normal and low-pass filtered speech. These patterns are similar to the normal and low-pass filtered speech using the Railways passage as

| | Span M-L | | | | | |
|---|---|---|---|---|---|---|
| | Normal Speech | | | Filtered Speech | | |
| Feature | Hertz | ERB | Semitone | Hertz | ERB | Semitone |
| confident | **0.485** | **0.532** | **0.556** √ | **0.421** | **0.499** √ | **0.491** |
| tense | **−0.425** | **−0.597** | **−0.714** √ | **−0.522** | **−0.612** | **−0.663** √ |
| harsh | -0.025 | -0.135 | -0.242 | **−0.385** | **−0.399** √ | **−0.365** |
| expressive | **0.363** | **0.473** | **0.520** √ | **0.635** | **0.741** √ | **0.714** |
| deep | -0.217 | -0.012 | **0.308** √ | **−0.462** √ | -0.202 | 0.140 |
| weak | **−0.436** | **−0.568** | **−0.663** √ | **−0.422** | **−0.462** | **−0.487** √ |
| irritated | **−0.352** | **−0.377** √ | **−0.360** | **−0.528** √ | **−0.512** | **−0.477** |
| happy | **0.493** | **0.548** √ | **0.544** | **0.366** | **0.528** | **0.612** √ |
| afraid | -0.273 | **−0.368** | **−0.470** √ | **−0.402** | **−0.407** √ | **−0.405** |
| relaxed | **0.585** | **0.670** | **0.726** √ | **0.341** | **0.511** | **0.614** √ |
| emphatic | **0.554** | **0.638** √ | **0.634** | **0.545** | **0.664** | **0.676** √ |
| bored | **−0.492** | **−0.583** | **−0.591** √ | **−0.497** | **−0.500** √ | **−0.360** |

**Table 4.8.** Comparison of the correlation coefficients between the M-L span, using the 3 different measurement scales, and listener judges' ratings of the Railways passage for 12 adjectives for both normal and filtered speech.

shown by the results in table 4.8.

Tables 4.9 to 4.12 present the results of similar analyses based on LTD measures of pitch range. It is clear from these analyses that a linguistically motivated measure of pitch range is far more successful in capturing the differences between listener judges' ratings of speakers than any of the widely used span measures based on long term distributional properties of f0. In many cases the measures based on long term distributional properties do not come close to showing significant relationships with speaker characteristics. An example of the correlation results is shown graphically in figure 4.3 which plots the average rating for *expressive* for each speaker with the M-L span measure on a semitone scale for each speaker (using open circles) and with a measure based on the 95th percentile minus the 5th percentile of all f0 for each speaker (using crosses), a measure used by Horii (1975). There is a clear positive correlation between width of span and more positivity of judgements of how expressive a speaker sounds using the linguistic measure of span. Such a clear correlation is not apparent using one of the usual statistical measures of range.

Given the assumption that positive characteristics are generally associated with wider spans and negative characteristics are generally associated with narrower spans, as has been shown both in previous studies (cf. Uldall 1960) and in the results of chapter 3, one correlation coefficient clearly stands out as being spurious in table 4.12. It is not considered likely that any true measure of span would correlate negatively with the feature *relaxed*. The long term distributional measures of level do make a reasonable estimation of the level characteristic, but again the results clearly show that the linguistically motivated measure F shows more and stronger correlations with speaker characteristics.

| Normal Speech | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Level (Hertz) | | | Span (Semitones) | | | |
| Feature | meanf0 | medianf0 | F | ± 2sds mean | 90% Range | 80% Range | M-L |
| confident | -0.036 | -0.076 | -0.279 | **0.333** | 0.004 | 0.089 | **0.529** √ |
| tense | 0.043 | 0.110 | 0.274 | -0.176 | 0.054 | -0.222 | **−0.410** √ |
| harsh | 0.234 | **0.365** √ | 0.184 | 0.189 | 0.229 | 0.132 | 0.114 |
| expressive | -0.080 | -0.127 | -0.212 | **0.414** √ | 0.076 | 0.087 | **0.413** |
| deep | **−0.821** | **−0.842** √ | **−0.834** | 0.276 | -0.290 | 0.170 | **0.333** √ |
| weak | 0.074 | 0.127 | 0.241 | **−0.499** | -0.039 | **−0.454** | **−0.632** √ |
| irritated | 0.134 | 0.207 | 0.294 | **−0.317** | 0.222 | 0.014 | **−0.546** √ |
| happy | 0.161 | 0.073 | 0.029 | 0.150 | -0.132 | -0.129 | 0.237 |
| afraid | 0.111 | 0.080 | 0.074 | -0.285 | **−0.347** | **−0.374** √ | -0.160 |
| relaxed | -0.239 | -0.295 | **−0.309** √ | 0.234 | -0.012 | 0.027 | **0.341** √ |
| emphatic | 0.098 | 0.065 | -0.045 | **0.428** | -0.014 | 0.150 | **0.465** √ |
| bored | 0.183 | 0.196 | 0.255 | -0.274 | -0.014 | -0.232 | **−0.453** √ |

**Table 4.9.** Comparison of the correlation coefficients between 2 level measures based on long term distributional properties and 1 level measure based on linguistic properties and a comparison of correlation coefficients between 3 span measures based on long term distributional properties and 1 span measure based on linguistic properties. Results in this table are based on the MTV passage and for normal speech.

### 4.5.2 Secondary analyses

The results of the primary analyses shows that the M-L measure of span, measured in semitones, and the F measure of level, measured in Hertz, are the most successful at capturing the differences in the perception of speaker characteristics. The results

| Normal Speech | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Level (Hertz) | | | | Span (Semitones) | | | | |
| Feature | meanf0 | medianf0 | F | | ± 2sds mean | | 90% Range | 80% Range | M-L | |
| confident | -0.068 | -0.125 | **−0.317** | √ | **0.512** | | 0.154 | 0.178 | **0.556** | √ |
| tense | 0.291 | **0.369** | **0.533** | √ | **−0.663** | | -0.031 | **−0.334** | **−0.714** | √ |
| harsh | **0.414** | **0.437** | √ **0.356** | | -0.188 | | 0.099 | 0.026 | -0.242 | |
| expressive | -0.213 | -0.274 | **0.360** | √ | **0.546** | √ | 0.063 | 0.289 | **0.520** | |
| deep | **−0.726** | **−0.733** | **−0.745** | √ | 0.195 | | -0.213 | 0.021 | **0.308** | √ |
| weak | 0.237 | **0.317** | **0.518** | √ | **−0.551** | | -0.083 | **−0.382** | **−0.663** | √ |
| irritated | 0.066 | 0.148 | 0.239 | | **−0.388** | √ | -0.083 | -0.100 | **−0.360** | |
| happy | 0.014 | -0.087 | -0.214 | | **0.462** | | -0.100 | 0.221 | **0.544** | √ |
| afraid | **0.356** | **0.385** | **0.536** | √ | **−0.470** | = | -0.035 | -0.228 | **−0.470** | = |
| relaxed | -0.084 | -0.151 | **−0.340** | √ | **0.716** | | 0.156 | **0.445** | **0.726** | √ |
| emphatic | -0.147 | -0.190 | **−0.384** | √ | **0.550** | | 0.068 | **0.321** | **0.634** | √ |
| bored | 0.048 | 0.119 | 0.264 | | **−0.525** | | -0.068 | **−0.304** | **−0.591** | √ |

**Table 4.10.** Comparison of the correlation coefficients between 2 level measures based on long term distributional properties and 1 level measure based on linguistic properties and a comparison of correlation coefficients between 3 span measures based on long term distributional properties and 1 span measure based on linguistic properties. Results in this table are based on the Railways passage and for normal speech.

| Filtered Speech | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Level (Hertz) | | | | Span (Semitones) | | | | |
| Feature | meanf0 | medianf0 | F | | ± 2sds mean | | 90% Range | 80% Range | M-L | |
| confident | -0.144 | -0.253 | -0.287 | | 0.047 | | -0.065 | -0.044 | **0.486** | √ |
| tense | **0.398** | **0.438** | √ **0.408** | | 0.173 | | **0.338** | √ 0.094 | -0.113 | |
| harsh | -0.058 | -0.082 | -0.117 | | 0.012 | | 0.058 | -0.016 | 0.076 | |
| expressive | 0.042 | -0.036 | **−0.297** | √ | **0.316** | | 0.176 | **0.404** | **0.647** | √ |
| deep | **−0.737** | √ **−0.730** | **−0.681** | | **−0.697** | | **−0.707** | √ **−0.531** | **−0.311** | |
| weak | **0.361** | **0.399** | √ **0.386** | | 0.221 | | **0.347** | √ 0.174 | -0.007 | |
| irritated | 0.038 | 0.092 | 0.211 | | -0.030 | | -0.024 | -0.082 | -0.109 | |
| happy | 0.026 | -0.054 | -0.235 | | 0.229 | | 0.125 | 0.194 | **0.614** | √ |
| afraid | **0.450** | **0.486** | √ **0.436** | | 0.249 | | **0.389** | √ 0.187 | -0.150 | |
| relaxed | **−0.476** | **−0.540** | **−0.589** | √ | -0.198 | | **−0.344** | √ -0.021 | 0.178 | |
| emphatic | -0.071 | -0.127 | **−0.360** | √ | 0.090 | | 0.063 | 0.177 | **0.384** | √ |
| bored | -0.183 | -0.147 | 0.049 | | -0.229 | | -0.288 | -0.197 | -0.288 | |

**Table 4.11.** Comparison of the correlation coefficients between 2 level measures based on long term distributional properties and 1 level measure based on linguistic properties and a comparison of correlation coefficients between 3 span measures based on long term distributional properties and 1 span measure based on linguistic properties. Results in this table are based on the MTV passage and for low-pass filtered speech.

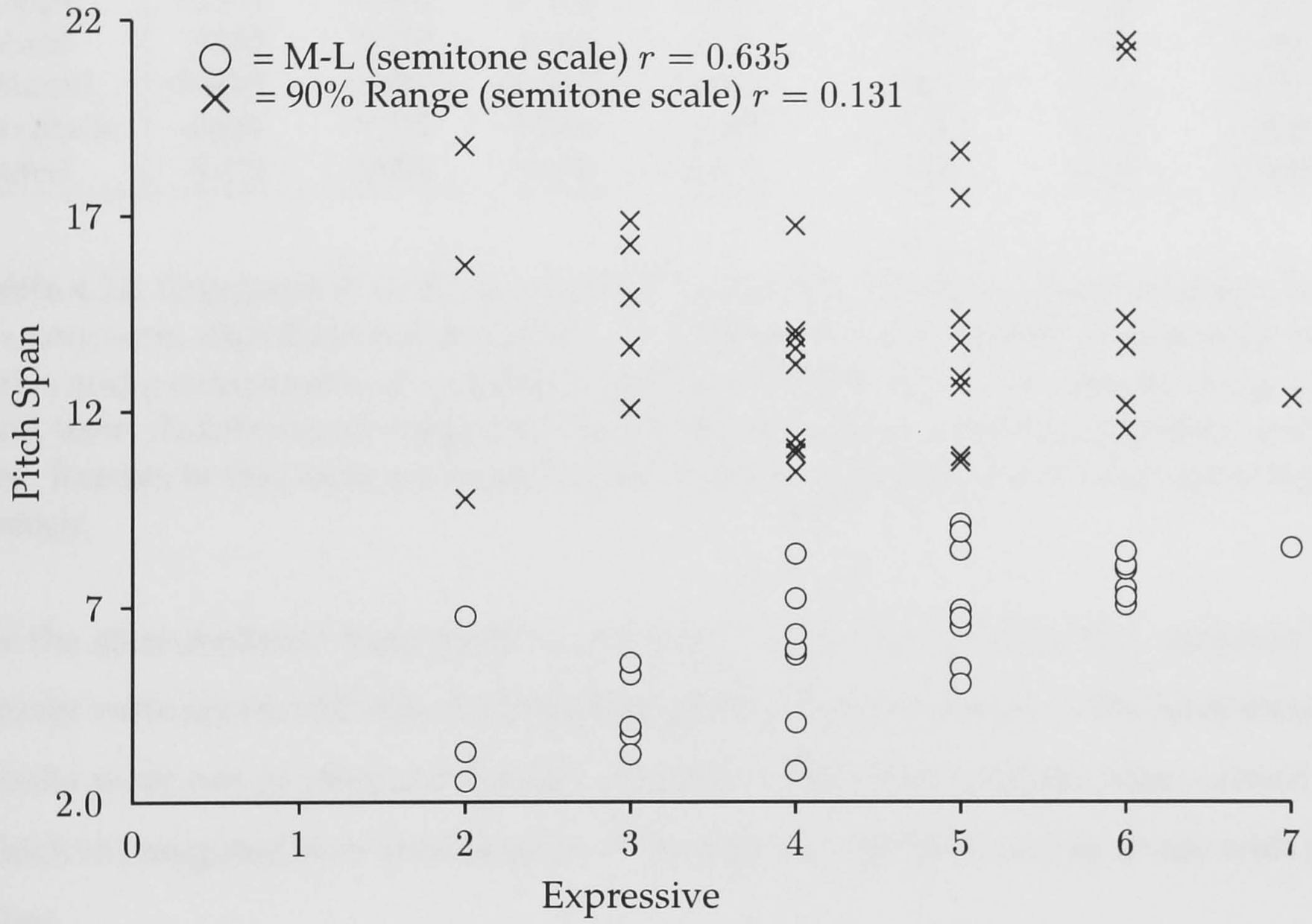



**Figure 4.3.** A scattergraph comparing a linguistic measure of span with a regularly used measure of span using general distributional properties of f0 against listener judges' ratings of 32 speakers on the characteristic "expressive"

| Filtered Speech | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Level (Hertz) | | | | Span (Semitones) | | | | |
| Feature | meanf0 | medianf0 | F | | ± 2sds mean | 90% Range | 80% Range | M-L | |
| confident | -0.148 | -0.219 | **−0.301** | √ | 0.087 | 0.017 | 0.046 | **0.421** | √ |
| tense | 0.187 | 0.257 | **0.369** | √ | -0.033 | 0.083 | -0.028 | **−0.522** | √ |
| harsh | -0.134 | -0.106 | 0.022 | | -0.209 | -0.200 | -0.232 | **−0.385** | √ |
| expressive | -0.007 | -0.118 | -0.235 | | 0.288 | 0.131 | **0.311** | **0.635** | √ |
| deep | **−0.882** | √ **−0.868** | **−0.737** | | **−0.809** | √ **−0.790** | **−0.610** | **−0.462** | |
| weak | 0.044 | 0.122 | **0.313** | √ | -0.028 | 0.038 | 0.064 | **−0.422** | √ |
| irritated | -0.094 | -0.021 | 0.015 | | -0.196 | -0.096 | -0.061 | **−0.528** | √ |
| happy | -0.170 | -0.264 | **−0.364** | √ | 0.009 | -0.072 | 0.063 | **0.366** | √ |
| afraid | 0.103 | 0.177 | 0.253 | | 0.015 | 0.068 | 0.149 | **−0.402** | √ |
| relaxed | **−0.417** | **−0.486** = | **−0.486** | = | -0.191 | **−0.349** | ? -0.094 | **0.341** | √ |
| emphatic | -0.086 | -0.172 | **−0.347** | √ | 0.195 | 0.111 | 0.212 | **0.545** | √ |
| bored | -0.079 | -0.005 | 0.050 | | **−0.338** | -0.260 | -0.262 | **−0.497** | √ |

**Table 4.12.** Comparison of the correlation coefficients between 2 level measures based on long term distributional properties and 1 level measure based on linguistic properties and a comparison of correlation coefficients between 3 span measures based on long term distributional properties and 1 span measure based on linguistic properties. Results in this table are based on the Railways passage and for low-pass filtered speech.

for the span measure were more convincing in the sense that the M-L measure was clearly more successful than the long term distributional measures. The level measure results were not so clear cut though. Further correlation analyses were carried out which investigated how similar each of the measures of level and span are with each other.

| | Hertz | | | ERB | | |
|---|---|---|---|---|---|---|
| | F | meanf0 | median f0 | F | mean | median f0 |
| L | **0.979** | **0.975** | **0.986** | **0.979** | **0.978** | **0.987** |
| F | | **0.947** | **0.957** | | **0.950** | **0.958** |
| mean f0 | | | **0.996** | | | **0.996** |

**Table 4.13.** Correlations between variations in measuring level for both Hertz and ERB measures. All $r$ coefficients that reach at least a significance level of $p < 0.05$ are in bold.

Table 4.13 shows the results of correlation analyses between the level measures L, F,

mean f0 and median f0 measured in both Hertz and ERB. It is clear from these results why it is very difficult to choose which method of measuring level is the strongest. All level measures show highly significant correlations with each other.

| | Hertz | | | | | |
|---|---|---|---|---|---|---|
| | H-F | M-L | M-F | meanf0±2sds | 95 - 5% f0 | 90 - 10% f0 |
| H-L | **0.936** | **0.905** | **0.773** | **0.753** | 0.075 | **0.305** |
| H-F | | **0.930** | **0.929** | **0.812** | 0.226 | **0.450** |
| M-L | | | **0.924** | **0.750** | 0.065 | **0.402** |
| M-F | | | | **0.759** | 0.243 | **0.521** |
| meanf0±2sds | | | | | **0.546** | **0.761** |
| 95 - 5% f0 | | | | | | **0.556** |

**Table 4.14.** Correlations between variations in measuring span for the semitone measure. All $r$ coefficients that reach at least a significance level of $p < 0.05$ are in bold.

Table 4.14 shows the results of correlation analyses between the span measures H-L, H-F, M-L, M-F, mean f0±2sds, 95 - 5% f0 and 90 - 10% f0 measured in semitones. The results in table 4.14 show that the 4 linguistic measures strongly correlate with each other. The correlations are clearly not as strong within the long term distributional measures group nor are the correlations between the linguistic measures and the long term distributional measures so strong.

| | M-L | |
|---|---|---|
| | ERB | semitones |
| Hz | **0.935** | **0.672** |
| ERB | | **0.890** |

**Table 4.15.** Correlations between variations in measuring M-L for the three measures Hertz, ERB and semitones. All $r$ coefficients that reach at least a significance level of $p < 0.05$ are in bold.

Table 4.15 shows the results of correlation analyses between the three different scales of measurement for the same span measurement; in this instance the M-L measure. The correlation coefficients show that changes in scale of measurement clearly alter the rankings of speakers with respect to the width of speaker span. This shows the

importance of trying to establish which scale of measurement best captures the difference in span across speakers. Results in the primary analyses for this experiment have shown that the logarithmic semitone scale is the most successful.

Correlation analyses have proved successful for the purposes of the current experiment to support the importance of choosing a suitable measure of span and level to characterise cross-speaker variation. The analyses, while successful, can be considered somewhat crude, in the sense that a lot of data is not utilised. To reduce all the listener judges' scores to a mode score makes a huge amount of data redundant. To address this problem a number of multiple regression analyses were carried out on the full data set of listener judges' scores for each speaker and for each feature. Table 4.16 shows the results of multiple regression analyses (adjusted $R^2$, total degrees of freedom, standardised Beta and p values) in which each speaker characteristic feature was the dependent variable, and span and level were the predictors. Generally speaking the results in table 4.16 show that the best linguistic measure of span and level accounts for more variation in the listener judges' scores than the most successful long term distributional properties measure of level and span, by having larger adjusted $R^2$ scores. Results also show that of the variation that span and level can account for in listener judges' scores, span is generally a much stronger predictor of that variation than level, except for the feature *deep*. It is clearly expected that level should be strongly related to the feature *deep*. Another point to make about the results, which show that neither level nor span can account for a large percentage of variation within the listener judges' responses, is that pitch range is only one of many possible carriers for communicating speaker characteristics, along with voice quality, intensity and durational effects.

Many possible sources of variation were controlled for within the current experiment. The Kruskal Wallis Test was run on the listener judges' responses to see if any of the controlled variables had any effect on those responses. Table 4.17 shows that for both the normal and filtered speech there were many significant effects of passage. The results show that, for the Railways passage, speakers were rated as being more

| feature | linguistic measure ML (semitones) and F (Hz) | | | | long term distributional measure meanf0±2sds (semitones) and meanf0 (Hz) | | | |
|---|---|---|---|---|---|---|---|---|
| | adjusted $R^2$ | total df | standardised Beta | p | adjusted $R^2$ | total df | standardised Beta | p |
| **confident** | 0.173 | 1535 | | | 0.103 | 1535 | | |
| span | | | 0.432 | 0.000 | | | 0.559 | 0.000 |
| level | | | 0.031 | 0.243 | | | -0.559 | 0.000 |
| **tense** | 0.129 | 1535 | | | 0.092 | 1535 | | |
| span | | | -0.310 | 0.000 | | | -0.447 | 0.000 |
| level | | | 0.087 | 0.002 | | | 0.551 | 0.000 |
| **harsh** | 0.001 | 1535 | | | 0.002 | 1535 | | |
| span | | | -0.006 | 0.828 | | | 0.038 | 0.409 |
| level | | | 0.047 | 0.112 | | | 0.020 | 0.664 |
| **expressive** | 0.184 | 1535 | | | 0.092 | 1535 | | |
| span | | | 0.482 | 0.000 | | | 0.554 | 0.000 |
| level | | | 0.141 | 0.000 | | | -0.470 | 0.000 |
| **deep** | 0.329 | 1535 | | | 0.324 | 1535 | | |
| span | | | -0.024 | 0.308 | | | 0.048 | 0.208 |
| level | | | -0.586 | 0.000 | | | -0.609 | 0.000 |
| **weak** | 0.181 | 1535 | | | 0.114 | 1535 | | |
| span | | | -0.380 | 0.000 | | | -0.505 | 0.000 |
| level | | | 0.082 | 0.002 | | | 0.616 | 0.000 |
| **irritated** | 0.028 | 1535 | | | 0.007 | 1535 | | |
| span | | | -0.181 | 0.000 | | | -0.152 | 0.001 |
| level | | | -0.023 | 0.425 | | | 0.161 | 0.000 |
| **happy** | 0.093 | 1535 | | | 0.052 | 1535 | | |
| span | | | 0.339 | 0.000 | | | 0.416 | 0.000 |
| level | | | 0.084 | 0.003 | | | -0.371 | 0.000 |
| **afraid** | 0.168 | 1535 | | | 0.111 | 1535 | | |
| span | | | -0.369 | 0.000 | | | -0.507 | 0.000 |
| level | | | 0.075 | 0.005 | | | 0.606 | 0.000 |
| **relaxed** | 0.146 | 1535 | | | 0.106 | 1535 | | |
| span | | | 0.331 | 0.000 | | | 0.477 | 0.000 |
| level | | | -0.090 | 0.001 | | | -0.591 | 0.000 |
| **emphatic** | 0.166 | 1535 | | | 0.089 | 1535 | | |
| span | | | 0.452 | 0.000 | | | 0.542 | 0.000 |
| level | | | 0.114 | 0.000 | | | -0.480 | 0.000 |
| **bored** | 0.096 | 1535 | | | 0.040 | 1535 | | |
| span | | | -0.358 | 0.000 | | | -0.360 | 0.000 |
| level | | | -0.153 | 0.000 | | | 0.261 | 0.000 |

**Table 4.16.** A comparison of results of multiple regression using the results for listeners' judgements of speaker characteristic features as the dependent variable and either a linguistic or a long term distributional measure for both span and level as the two independent variables.

| feature | normal speech | | | low pass filtered speech | | |
|---|---|---|---|---|---|---|
| | MTV | $\chi^2$ | Railways | MTV | $\chi^2$ | Railways |
| confident | | **17.656** | √ | | **9.566** | √ |
| tense | √ | **18.683** | | √ | **19.774** | |
| harsh | √ | **5.222** | | √ | **6.915** | |
| expressive | | **35.339** | √ | | **12.102** | √ |
| deep | | 0.527 | | | 0.776 | |
| weak | √ | **8.057** | | √ | **4.662** | |
| irritated | | 3.441 | | √ | **18.625** | |
| happy | | **27.903** | √ | | **24.246** | √ |
| afraid | √ | **9.234** | | √ | **11.539** | |
| relaxed | | **24.662** | √ | | **17.402** | √ |
| emphatic | | **26.456** | √ | | **13.727** | √ |
| bored | √ | **13.811** | | √ | **20.402** | |

**Table 4.17.** Results of the Kruskal Wallis Test showing effects of passage on listeners judgements of speakers on 12 speaker characteristics for normal speech and filtered speech. $\chi^2$ that that are significant to at least the $p < 0.05$ level are indicated in bold type. √ indicates which passage gets a higher rating for each feature for which there is a significant difference.

*confident, expressive happy, relaxed* and *emphatic* than for the MTV passage. For the MTV passage, speakers were rated as being more *tense, harsh, weak, afraid* and *bored* than for the Railways passage. Results in table 4.18 show that for both the MTV passage and the Railways passage, the filtered speech of speakers was rated as more *harsh* and *deep* than for the unfiltered speech. For the normal speech, speakers were rated as being more *confident* and *happy* than for the filtered speech. These results show that it was important to report results of correlation analyses for both passages and both filtering types separately.

Results in table 4.19 report on the effects of speaker sex on listener judges' responses. Results show that men are rated more positively by both male and female listeners. Female listeners judged male speakers to be more *confident, deep, happy* and *relaxed* than female speakers and judged female speakers to be more *tense, harsh, weak, irritated* and *afraid* than the male speakers. These results were similar for the pattern in male listeners' responses.

| feature | MTV | | | Railways | | |
|---|---|---|---|---|---|---|
| | normal | $\chi^2$ | filtered | normal | $\chi^2$ | filtered |
| confident | √ | **5.803** | | √ | **17.768** | |
| tense | | 1.515 | | | 0.420 | |
| harsh | | **9.139** | √ | | **8.705** | √ |
| expressive | | 0.054 | | √ | **6.067** | |
| deep | | **11.761** | √ | | **11.458** | √ |
| weak | | 0.027 | | | 1.088 | |
| irritated | | **21.181** | √ | | 3.690 | |
| happy | √ | **3.986** | | √ | **7.344** | |
| afraid | | 0.404 | | | 0.117 | |
| relaxed | | 3.229 | | | 0.068 | |
| emphatic | | 0.471 | | | 1.075 | |
| bored | | **4.218** | √ | | 0.930 | |

**Table 4.18.** Results of the Kruskal Wallis Test showing effects of filtering on listeners judgements of speakers on 12 speaker characteristics for the MTV passage and Railways passage. $\chi^2$ that are significant to at least the $p < 0.05$ level are indicated in bold type. √ indicates which filtering type gets a higher rating for each feature for which there is a significant difference.

| feature | male listeners | | | female listeners | | |
|---|---|---|---|---|---|---|
| | male speakers | $\chi^2$ | female speakers | male speakers | $\chi^2$ | female speakers |
| confident | | 1.607 | | √ | **11.171** | |
| tense | | **18.908** | √ | | **31.511** | √ |
| harsh | | **8.313** | √ | | **5.428** | √ |
| expressive | | 0.835 | | | 0.470 | |
| deep | √ | **116.841** | | √ | **428.445** | |
| weak | | **9.986** | √ | | **40.339** | √ |
| irritated | | **4.480** | √ | | **5.847** | √ |
| happy | | 2.694 | | √ | **4.019** | |
| afraid | | **5.340** | √ | | **50.044** | √ |
| relaxed | √ | **17.329** | | √ | **49.487** | |
| emphatic | | 2.349 | | | 2.543 | |
| bored | | 2.782 | | | 1.108 | |

**Table 4.19.** Results of the Kruskal Wallis Test showing effects of sex of speaker on listeners judgements of speakers on 12 speaker characteristics for male speakers and female listeners. $\chi^2$ that that are significant to at least the $p < 0.05$ level are indicated in bold type. √ indicates which speaker sex gets a higher rating for each feature for which there is a significant difference.

| feature | Scottish Speakers | | | English Speakers | | |
|---|---|---|---|---|---|---|
| | Scottish listeners | $\chi^2$ | English listeners | Scottish listeners | $\chi^2$ | English listeners |
| confident | | 0.261 | | | 0.857 | |
| tense | | **6.689** | √ | | **14.434** | √ |
| harsh | | 1.796 | | | 3.825 | |
| expressive | | 0.857 | | | 0.041 | |
| deep | | 0.821 | | | 1.686 | |
| weak | | 3.732 | | | 3.578 | |
| irritated | | 0.645 | | | **4.464** | √ |
| happy | | 3.669 | | | 1.976 | |
| afraid | | **10.959** | √ | | **6.249** | √ |
| relaxed | | **4.777** | √ | | 0.048 | |
| emphatic | | **5.292** | √ | | **3.941** | √ |
| bored | | 2.585 | | | **5.246** | √ |

**Table 4.20.** Results of the Kruskal Wallis Test showing effects of listener accent on listeners judgements of speakers on 12 speaker characteristics for Scottish speakers and English speakers. $\chi^2$ that that are significant to at least the $p < 0.05$ level are indicated in bold type. √ indicates which listener accent gets a higher rating for each feature for which there is a significant difference.

Results in table 4.20 report on the effects of listener accent on listener judges' responses. Results show that English listeners were using more extreme scores for some of the features for judging both Scottish and English speakers, though there does not seem to be any pattern as to whether these features themselves are necessarily negative or positive. English listeners judged Scottish speakers to be more *tense, afraid, relaxed* and *emphatic* than Scottish listeners and judged English speakers to be more *tense, irritated, afraid, emphatic* and *bored* than the Scottish listeners. From the results of the current study it would be difficult to claim that there is a minority group reaction or a reflection of the influence of community-wide stereotypes of English and Scottish accents as suggested by Lambert (1972a), or any sense of accent loyality suggested in Cheyne (1970). The Cheyne (1970) experiment was run in Glasgow when Scottish listeners were involved, and in London when English listeners were involved. The lack of similarity of the current experimental results, in patterning to previous results, could well be based on the fact that all the subjects were attending a Scottish University. Listeners simply may not hold the same stereotypes as they all choose to favour

study in Scotland as opposed to England, or perhaps may feel very aware of following stereotypical behaviour and try to rate speakers more favourably where possible.

## 4.6 Conclusions and Discussion

From the pitch data collected in experiment 2, 4 possible dimensions of span (H-L, H-F, M-L and M-F) and two measures of level (L and F) were investigated. All measures of level and span for 32 speakers were correlated with listener judgements of certain speaker characteristics of those 32 speakers. 4 sets of listener judgements were used; results for each passage and each type of speech presentation, being MTV passage normal speech, Railways passage normal speech, MTV passage low pass filtered speech and Railways passage low pass filtered speech.

As in experiment 1, it is assumed the best measures of level and span are those that show the strongest correlations with the listener judges' data. For overall span the strongest correlations were with the M-L span. Results for level were similar for both L and F, but F is considered the most successful in the current experiment.

Span measures were measured on three different scales; a linear scale measured in Hertz, a logarithmic scale measured in musical semitones and a scale based on the human auditory system, the ERB-rate scale. Again using correlation results with the listener judges' data, no differences were found if level is represented using the Hertz scale or the ERB scale. Hertz was selected as the better measurement unit for level, based purely on the simplicity of measurement. It was shown that musical semitones is the scale which best represents variation in span across speakers.

Correlation results of the M-L span, measured using semitones, were compared to the correlation results of 3 previously used measures of span based on long term distributional properties of f0. These long term distributional measures showed very few significant correlations with the listeners judges' data. Across all the data, the M-L

measure based on tonal targets in speech, showed stronger correlations with listener judges' data compared to any competing measures.

Correlation results of the F level measure in Hertz were compared to the correlation results of 2 previously used measures of level based on long term distributional properties of f0. Results show that the correlation results for all measures of level were very similar. This was supported by results of actually correlating all the measures of level with each other. All suggested measures strongly correlated with each other.

The central claim of this thesis has been that tonal targets, such as those suggested in Shriberg *et al.* (1996), not only capture differences in pitch range across speakers successfully, but do so more successfully than any other measure. Results of experiment 2 clearly support this claim. One possible explanation for the success of the M-L measure of span compared to the various statistical measures could be to do with the exclusion of the uncharacteristically high first accent which will always skew the distribution. Comparing the results in table F.3 with table F.9 and table F.4 with table F.10 in appendix F, in the vast majority of cases the H-L measure (which of course includes the sentence initial high accent) out-performs the best of the long term distributional measures. As this is not as clear cut as the difference between the M-L measure and the long term distributional models we certainly will not write off the possibility of the importance of the sentence initial accent.

Results of secondary analyses confirm that the decision to separate the data into separate passages (Railways and MTV) and separate speech types (normal and low-pass filtered) was a necessary requirement. There are further arguments to suggest that seperate analyses should have been performed between the different sexes for both listeners and speakers and to a lesser extent for accent type as well. Two reasons are provided for why this was not done. Firstly, one of the key objectives of the current study was that it should be relatively large scale. Given the resources available, and given limits on the length of time a subject can be expected to concentrate fully on a single experimental task, 32 speakers was considered the optimal number. Any less

than this would take away one of the key components of the experiment. Secondly, the results highlight an interesting question as to the different perceptions of male and female speakers. Given the fact that more positive characteristics are attributable to wider spans, are men attributed with more positive characteristics because they are men and listeners are just giving stereotypical judgements, or can the responses be attributed to span differences? If the speakers in experiment 2 are put in order of span width from narrow to wide using the semitone scale, the list, with respect to sex of speaker, would read (f, f, f, m, f, f, f, f, m, m, f, m, f, m, m, f, m, f, m, m, f, m, f, m, m, f, m, f, m, f, m, m), where "f" stands for female and "m" stands for male. From this list it can be seen that generally speaking the males used in experiment 2 fill the top end of the list and females tend to have narrower spans.

One of the key differences between a measure of span based on tonal targets in speech and those measures based on long term distributional properties is that the former places no special importance on the middle of the f0 distribution. Long term distributional measures make great use of the central tendency of f0 and assume that the topline and bottomline for span should be related at some point on equal sides of the central tendency, whether it be two standard deviations plus or minus around the mean, or the 95th and the 5th percentile. This assumes that f0 is distributed normally around the mean. The results of skew and kurtosis for each speaker in table A.7 in appendix A shows that f0 is not normally distributed around the mean. Shriberg *et al.* (1996) have already demonstrated that within speaker expansion in pitch range occurs bottom up with sentence final low being a relatively stable base, as opposed to radiating both upwards and downwards away from the mean. The Shriberg *et al.* (1996) model used tonal target points to successfully predict within speaker variation in pitch range expansion. The current experiment adds support to the linguistic modelling of pitch range, by showing that a measure of span and level using tonal targets also better captures variation in span across speakers.

Nevertheless future research should attempt to find properties of speakers' long term distributions of f0 that approximate the parameters used in the model. It will be of

great benefit for future research if such properties can be found, as the current method of data collection is very time consuming and labour intensive. As has been shown, such a measure will have to be based on more than just the basic elements of the long term distribution. Results from the current study have shown that pitch range is not satisfactorily approximated by the distribution of f0 around the mean. We suggest that one possible way to attempt a more satisfactory estimation of pitch range would be to somehow incorporate the further distributional properties of skew and kurtosis. Using skew and kurtosis measures would avoid the fallacy that f0 is distributed normally around the mean.

One of the key motivations for the two percentile measures of range (both the 80% and the 90% range measures) is that they assume that cutting out both the outer limits of the f0 distribution will cut off all the measurement errors that often occur in pitch extraction. It may be possible to use skew and kurtosis measures to set speaker-specific cut-off points. A possible suggestion is that a speaker with a strong negatively skewed distribution could be an indication that that speaker actually uses the lower end of his or her range more than the upper range. A negatively skewed distribution may also indicate that there are less errors in tracking the pitch at the lower end of the range and that the top end of the distribution is particularly messy in terms of pitch extraction. By using skew and kurtosis it might be possible to suggest that the example speaker's pitch range would best be estimated by taking the range between the 88th percentile and the 3rd percentile. Clearly this is just a suggestion as to how speakers' long term distribution of f0 could still be used to approximate the parameters used in the model suggested by the results of the current study, and these ideas should be followed up with further research.

# Chapter 5

# Experiment 3

## 5.1 Introduction

Chapter 5 reports on two studies combined in a single experiment. The first study uses resynthesised speech as a tool to further examine the span measures established in experiment 2 (chapter 4), which in turn were based on measurement targets established for Dutch speech in Ladd & Terken (1995) and Shriberg *et al.* (1996). For reference, we shall be calling this first study the "resynthesis" study. The second study in this chapter replicates to some extent experiment 2 and continues the central theme of this whole thesis, which is to examine a linguistic model of pitch range and compare such a model with previous suggested models of pitch range based on long term distrbutional properties of f0. For reference, we shall be calling the second study the "replication" study. The "replication" study is essentially a scaled down version of experiment 2: we use 8 speakers instead of 32 speakers, 1 passage instead of 2 passages and only the normal speech condition with no low pass filtered speech condition.

At the very beginning of this thesis, pitch span was defined in the simplest terms as the difference between some topline and some bottomline from all the f0 values used by a speaker (section 1.1). The results of experiment 2 have established that an

average of the non-sentence-intial peaks (M) is the best topline and the average of the post-accent valleys (L) is the best bottomline, with respect to capturing differences in the perception of various speaker characteristics.

In section 3.1 it was noted that pitch range has most strongly correlated with the evaluation dimension of a three dimensional model of emotion (Pakosz 1982). This result is supported by results of the multi-dimensional scaling analysis in chapter 3 (see figure 3.7). Results from experiment 2 (chapter 4) generally show that speakers characterised as being more positive (more confident, relaxed, happy) have wider pitch spans and speakers characterised as being more negative (more tense, irritated) have narrower pitch spans. Essentially, though, it seems that pitch range only has a very coarse grained effect on speaker characterisation.

The main purpose of the resynthesis experiment is to report on an investigation into a more detailed analysis of pitch span. A model of pitch span involving greater detail may allow us to make more specific claims about the effects of span on speaker characteristics. The standard "difference between some topline and some bottomline" model at best accounts for positive/negative judgements of speaker characteristics. A more detailed model may be able to tease apart variations between how, for example, a speaker may express being more confident as compared to being more emphatic.

In section 1.4.1 a definition of pitch range as "a global, or at least phrase-sized choice of pitch scaling parameters" (Liberman & Pierrehumbert 1984) was introduced. This definition was mentioned in the discussion of the reanalysis of the H*+H...H* sequence introduced by Beckman & Pierrehumbert (1986). In the 1986 analysis of the problematic tonal sequence under discussion, the H*+H...H* analysis was replaced with just two H* accents and the sustained high transition was explained as an effect of a local elevation of pitch level and compression of pitch span, as described in figure 1.4. One of the main objections against the Beckman & Pierrehumbert (1986) analysis is that the local manipulation of pitch range does not fit well with the idea that pitch range is a more global feature (Ladd 1996). If results of experiment 3 show that a more detailed

description of pitch range does offer further insight into the perception of speaker characteristics, it opens the possibility that fine-grained manipulations of pitch span can not only be used to express certain speaker characteristics but also can be used to explain linguistic phenomenon such as the reanalysis of the H*+H...H* sequence discussed above.

The current study looks into greater detail at the top of the speaker range. In experiment 2, two potential toplines were established: H, the average of a speaker's sentence-intial high, and M, the average of a speaker's non-sentence-initial high. These two toplines were essentially put into competition against each other to see which measure best captured listener judges' perception of speaker characteristics. Results showed that M was the most suitable top line. For the current experiment, resynthesis is used as a tool to investigate the potential that both H and M together may provide a more complete description of speaker span and provide necessary information for more detailed speaker characterisation. In the resynthesis experiment to be reported on, 4 versions of speakers' resynthesised voices were judged on a set of phonetic and pragmatic criteria. The 4 versions were "normal", "raised M", "raised "H" and "raised M and raised H". More details of these 4 versions are provided in section 5.2.2 below. One of the key elements of the design is to see if speakers are characterised in any more specific detail in the "raised M and raised H" condition compared to just the "raised M" condition.

Brown *et al.* (1973) reports on a similar resynthesis study in which pitch range was manipulated by increasing the variance of f0 from the mean by 50%. The increase span condition showed that speakers were judged as sounding more *benevolent*. There was also a trend (though not statistically significant) showing that speakers were judged as more *competent*. A further analysis in the current study is made to see if speakers are rated more postively in the fully increased pitch span condition ("raised M and raised H") than in the unchanged ("normal") span condition.

## 5.2 Stimuli Design and Analysis

### 5.2.1 Speakers

The speech of 4 Scottish males and 4 Scottish females was used for the "resynthesis" study. All 8 speakers were chosen from the original 32 speakers used in experiment 2. Speakers were selected so as to have a wide variety of both level and span. The ages of the 8 speakers in the "resynthesis" study are shown in table 4.1 in chapter 4.

The speech of 2 Scottish males, 2 Scottish females, 2 English males and 2 English females was used for the "replication" study. In the "replication" study there is again the assumption that there is enough variation in the voices to be able to make "ratable" differences in speaker characteristics, rather than relying on recordings of acted or simulated characteristics. All 8 speakers were taken from the English speech database collected for experiment 2, but had not been previously used in any other study. There was a mixed spread of ages used in the "replication" study. The age of each speaker is reported in table 5.2.

### 5.2.2 Speech Materials

For the "resynthesis" experiment the recordings of the MTV passage (see appendix C) were used. For the "replication" study, recordings of the Railways passage (also in appendix C) were used. For the purposes of the "replication" study, only normal speech was used; there was no low pass filtering condition.

**Resynthesis**

The Praat package, developed at the University of Amsterdam, was used for the manipulation of pitch span on the chosen speech files. Four different pitch span versions were resynthesised for each speaker. Praat uses the Pitch Synchronous Overlap and

Add (PSOLA) method for resynthesis.

Version 1 was a near normal version of the original speech file for each speaker. The only difference between version 1 and the original speech file was that version 1 was smoothed. Praat has a smoothing function which takes away all the octave error data from the original speech files. Version 2 was a resynthesised speech file in which the sentence-intial peaks (H) had been raised. Version 3 was a resynthesised speech file in which all the H peaks were unaltered from version 1, but all the non-sentence-initial peaks (M) had been raised. Version 4 was a resynthesised speech file in which all the H and M peaks had been raised. For clarification, the difference between the target points being raised in all four resynthesis versions are shown graphically in figure 5.1.

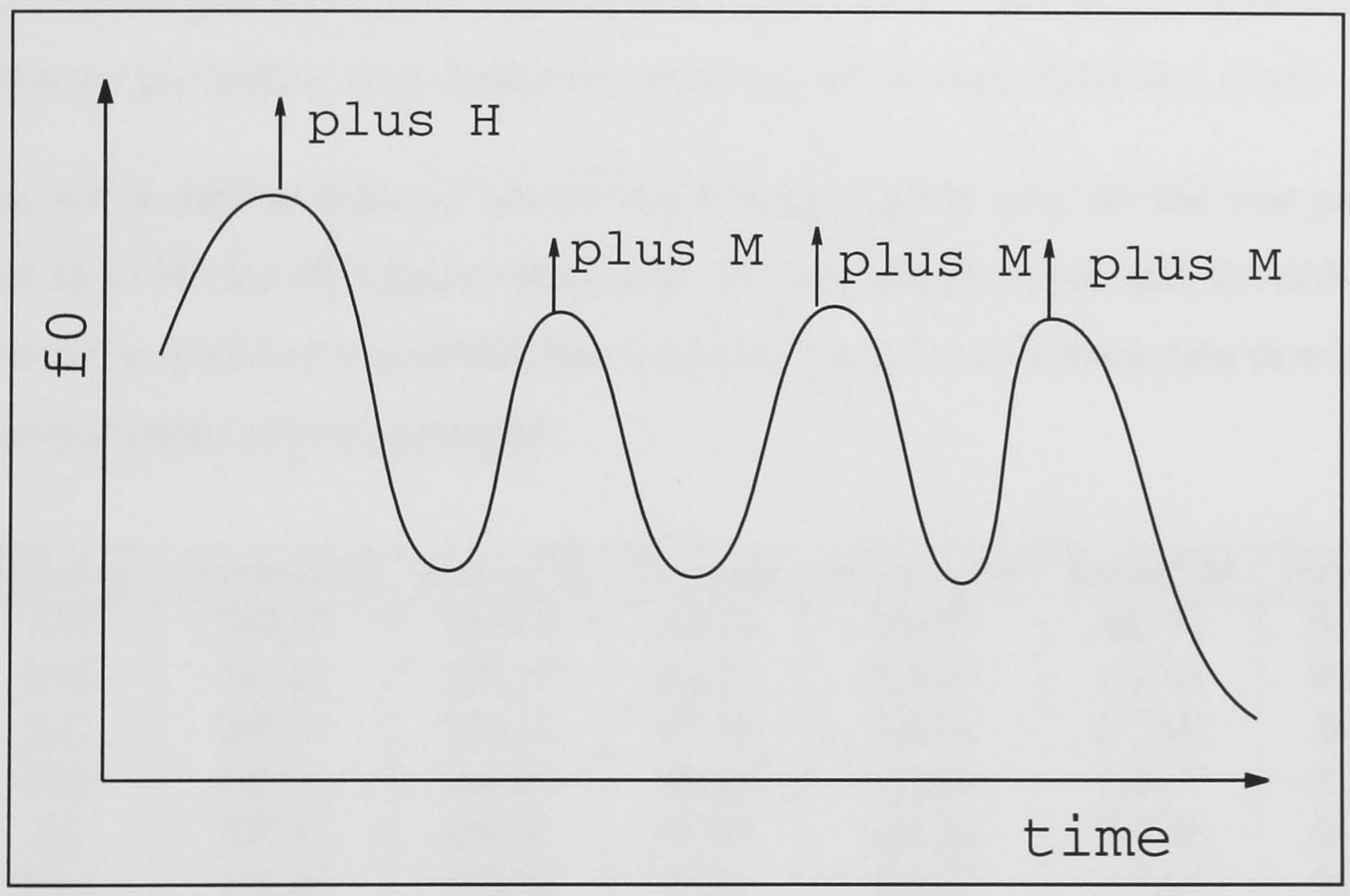



**Figure 5.1.** Locations for increases in span for the H and M parameters on an idealised speaker contour representing the normal "smoothed" version

The criterion used in deciding the extent of all the pitch span expansions was purely impressionistic. The spans for each speaker were widened by raising the pitch at the

selected tonal targets as much as possible while still making sure the speech continued to sound natural. From table 5.1 it can be seen that there is clear variation in the amount of span expansion for each speaker. For example AW normally has an average sentence-initial high of 339.67Hz and in the resynthesised versions that included raised H, this average was increased to 429Hz; an increase of 89.33Hz. This increase was roughly matched when increasing the other accent peaks (M) which were raised on average by 81.95Hz.

### 5.2.3 Pitch Range Analysis

Pitch range data was extracted from the resynthesised speech for the "resynthesis" experiment and from the normal speech used in the "replication" experiment using the same procedure as set out in experiment 2 (cf. sections 3.2.4 and 4.2.4).

The set of data in table 5.1 shows the averaged pitch data for the two positions H and M in Hz for all 8 speakers used in the resynthesis study and the difference between the speakers' smoothed data (version 1) and the increased data due to the span manipulations after resynthesis.

| Speaker | Normal H | Raised H | Increase | Normal M | Raised M | Increase |
|---|---|---|---|---|---|---|
| AW | 339.67 | 429.00 | 89.33 | 258.78 | 340.73 | 81.95 |
| KW | 163.67 | 229.17 | 65.50 | 136.80 | 187.43 | 50.63 |
| KG | 292.67 | 320.33 | 27.66 | 244.96 | 275.92 | 30.96 |
| TM | 147.67 | 184.33 | 36.66 | 115.65 | 136.77 | 21.12 |
| SS | 335.67 | 420.67 | 85.00 | 243.96 | 307.88 | 63.92 |
| HM | 191.00 | 258.33 | 67.33 | 149.35 | 209.08 | 59.73 |
| JC | 258.50 | 291.50 | 33.00 | 222.88 | 265.46 | 42.58 |
| JS | 192.00 | 246.50 | 54.50 | 153.77 | 185.64 | 31.87 |

**Table 5.1.** Variation in pitch measures due to resynthesis

The set of data shown in table 5.2 shows the averaged pitch range results for each speaker used in the "replication" study. Because of the results from experiment 2, the span measure for each speaker is represented using the logarithmic musical semitone

| | | | | Span (semitones) | | | | Level (Hz) | |
|---|---|---|---|---|---|---|---|---|---|
| Speaker | Age | Sex | Accent | H-F | H-L | M-F | M-L | F | L |
| FG | 24 | male | Scottish | 16.29 | 12.90 | 11.22 | 7.83 | 87.50 | 106.41 |
| KP | 20 | male | Scottish | 11.46 | 9.38 | 5.67 | 3.59 | 73.17 | 82.50 |
| AL | 63 | male | English | 15.51 | 12.33 | 9.58 | 6.40 | 70.83 | 85.13 |
| JM | 67 | male | English | 12.82 | 9.60 | 8.95 | 5.73 | 97.83 | 117.83 |
| CS | 32 | female | English | 16.68 | 12.38 | 11.97 | 7.67 | 124.67 | 159.78 |
| SP | 35 | female | English | 12.11 | 9.57 | 7.68 | 5.14 | 130.33 | 150.92 |
| JW | 38 | female | Scottish | 21.11 | 17.83 | 15.66 | 12.38 | 113.67 | 137.37 |
| LC | 47 | female | Scottish | 13.98 | 11.29 | 9.75 | 7.06 | 135.57 | 158.36 |

**Table 5.2.** Span and level measures for each speaker used in experiment 3, measured in Hz

scale, and the level is represented using the linear Hz scale. Figure 5.2 shows the variation of span and level across the 8 speakers that were used for the "replication" study. There is a clear divide between the men and the women in this graph. JW can be described as having a high level and a wide span, KP has both a low level and a narrow span and speaker SP has a high level and low span.

## 5.3 Perception Experiment

The perception experiment, which incorporated the speech for both the "replication" and "resynthesis" studies, follows the same methodology used for both experiment 1 and experiment 2. The aim of the perception experiment is to get profiles of speaker characteristics based on a selection of pragmatic and phonetic criteria for each speaker. This was achieved by asking subjects to fill in a rating form while listening to speech recordings of each speaker. From the listeners' responses it is possible to see how each speaker can be characterised.

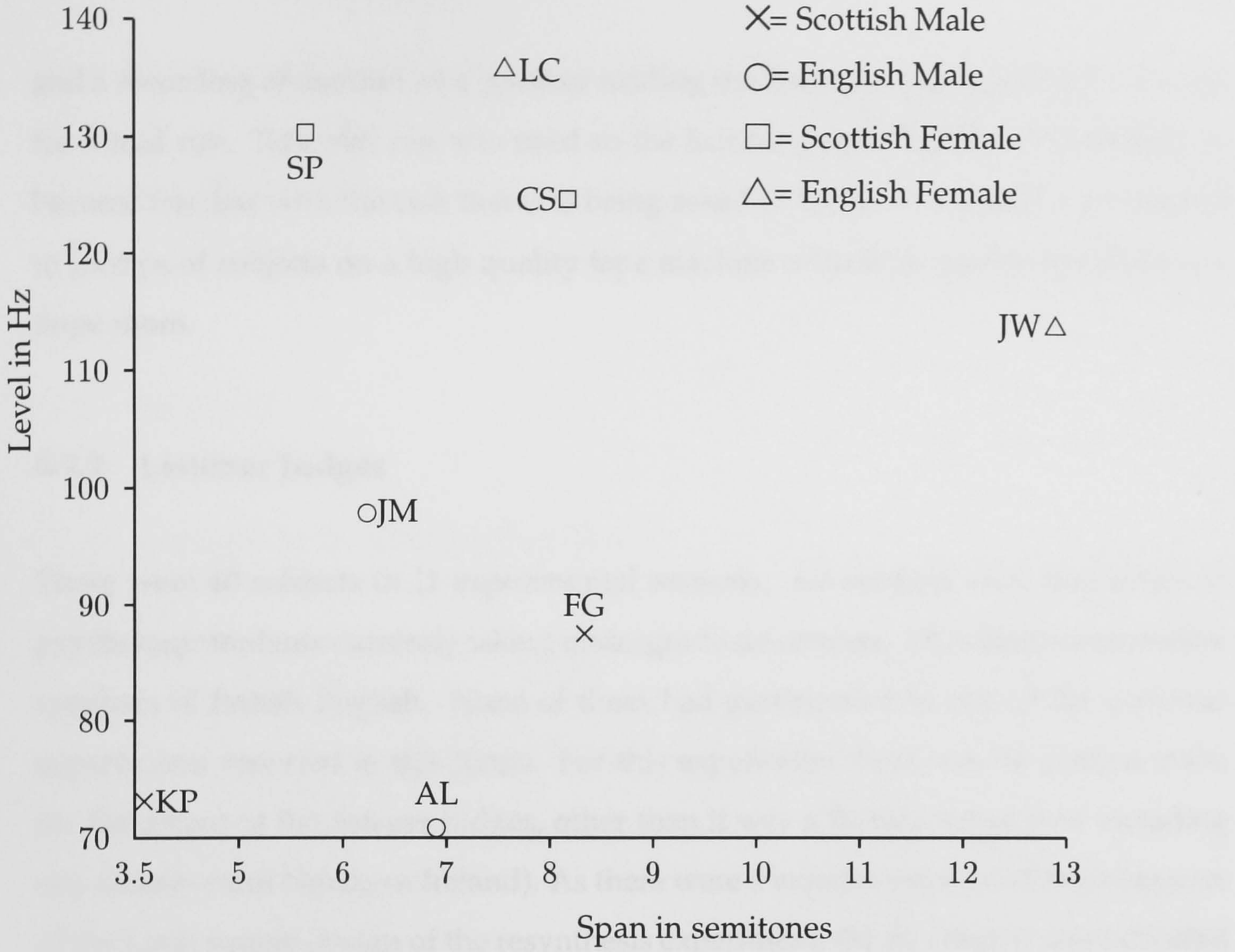



**Figure 5.2.** Span and Level of the eight speakers in the replication study

### 5.3.1 Speech Materials

Four experimental tapes were prepared. On each tape was all the speech material for a full experimental run. Each tape represented one of four different groups for the "resynthesis" experiment which was based on a Latin square design. In each experimental condition each speaker was only heard once and there were 2 examples of the 4 different versions of each passage (versions discussed in section 5.2.2). On each tape there was also the normal versions of the 8 speakers being used for the "replication" study. In total 16 speakers were being rated for the purposes of the two studies. For each tape the 16 speakers were put in different but random orders.

On each tape there was also a recording of one new speaker reading the MTV passage

and a recording of another new speaker reading the Railways passage that were used for a trial run. This trial run was used so the listener judges had the opportunity to become familiar with the task that was being asked of them. The stimuli were played to groups of subjects on a high quality tape machine with high quality speakers in a large room.

### 5.3.2 Listener Judges

There were 40 subjects in 11 experimental sessions. All subjects were linguistics or psychology students currently taking undergraduate courses. All subjects were native speakers of British English. None of them had participated in any of the previous experiments reported in this thesis. For this experiment there was no control made for the accent of the listener judges, other than it was a British accent (not including any accents from Northern Ireland). As there were 4 experimental conditions because of the Latin square design of the resynthesis experiment, the 40 subjects were divided up into 4 equal groups of 10.

### 5.3.3 Rating Form

For the current experiment there were 12 features to be judged - *confident, tense, harsh, expressive, deep, weak, irritated, happy, afraid,relaxed, emphatic* and *bored*. The same 7 point unipolar scale used in experiment 1 (chapter 3) was used for this experiment (cf. section 3.4.1).

### 5.3.4 Experimental session

Subjects were given a rating form booklet. On the cover of the booklet there were general instructions as to how the experiment would be run, as well as 2 examples of the rating form. Once they had finished reading the instructions, the subjects were

given an opportunity to ask questions. Then a practice run was carried out, using the speech of 2 different speakers, one reading the MTV passage and one reading the Railways passage. After a further opportunity for questions, the experiment proper was run. The whole experimental session lasted 25 minutes.

## 5.4 Results

### 5.4.1 Results of the Replication Study

| feature | Men | | | | Women | | | |
|---|---|---|---|---|---|---|---|---|
| | Scottish | | English | | | | Scottish | |
| | FG | KP | AL | JM | CS | SP | JW | LC |
| confident | 6 | 5 | 3 | 5 | 6 | 5 | 6 | 5 |
| tense | 2 | 3 | 3 | 3 | 2 | 3 | 2 | 3 |
| harsh | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 3 |
| expressive | 6 | 2 | 3 | 3 | 5 | 3 | 5 | 5 |
| deep | 5 | 6 | 5 | 2 | 2 | 3 | 2 | 2 |
| weak | 1 | 3 | 3 | 3 | 2 | 2 | 1 | 1 |
| irritated | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| happy | 4 | 2 | 2 | 2 | 3 | 2 | 5 | 2 |
| afraid | 1 | 3 | 3 | 3 | 1 | 1 | 1 | 1 |
| relaxed | 5 | 2 | 2 | 2 | 6 | 2 | 5 | 5 |
| emphatic | 5 | 2 | 2 | 2 | 6 | 3 | 5 | 4 |
| bored | 3 | 3 | 2 | 2 | 1 | 3 | 1 | 2 |

**Table 5.3.** Mode results of speech for all speakers reading the Railways passage for experiment 3

The *mode* for each feature for each speaker averaging across all listeners was calculated using the SPSS statistical package. The full results can be found in table 5.3. For example, the results in table 5.3 show that FG, a Scottish male, was judged as being "6" on the confident scale, so generally the listeners perceived FG to be very confident.

The relationship between the pitch range parameters for each speaker and the results of the perception study were established by calculating Spearman's rank correlation

| Feature | Level | | | Span | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | | F | M-L | | M-F | | H-L | | H-F | |
| confident | 0.391 | | 0.274 | **0.717** | = | **0.717** | = | **0.652** | | **0.652** | |
| tense | -0.282 | | -0.056 | **−0.845** | = | **−0.845** | = | **−0.845** | = | **−0.845** | = |
| harsh | 0.077 | | 0.309 | -0.463 | | -0.463 | | -0.617 | | -0.617 | |
| expressive | 0.476 | | 0.350 | **0.901** | √ | **0.851** | | **0.826** | | **0.776** | |
| deep | **−0.805** | √ | **−0.702** | -0.447 | | -0.549 | | -0.345 | | -0.447 | |
| weak | -0.504 | | -0.567 | **−0.756** | √ | **−0.693** | | **−0.630** | | -0.567 | |
| irritated | -0.504 | | -0.252 | **−0.630** | | **−0.756** | = | **−0.630** | | **−0.756** | = |
| happy | 0.191 | | 0.027 | **0.873** | = | **0.846** | | **0.873** | = | **0.846** | |
| afraid | **−0.732** | = | **−0.732** | -0.620 | | -0.620 | | -0.507 | | -0.507 | |
| relaxed | **0.639** | √ | 0.443 | **0.809** | | **0.861** | √ | **0.717** | | **0.769** | |
| emphatic | **0.663** | √ | 0.491 | **0.798** | | **0.835** | √ | **0.724** | | **0.761** | |
| bored | -0.517 | | -0.239 | -0.580 | | **−0.718** | = | -0.580 | | **−0.718** | = |

**Table 5.4.** Results of correlation analyses for 2 linguistic measures of (measured in Hz) level and 4 linguistic measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the Railways passage. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

coefficients ($\rho$). Table 5.4 shows the correlation coefficient results comparing the Railways passage mode results with the linguistic measures of level, measured in Hertz, and the linguistic measures of span, measured in semitones. Just these measures are featured in this study due to the results established in experiment 2. All $\rho$ that are significant ($p < 0.05$) are in a bold font. For example, table 5.4 shows that the features *deep, afraid, relaxed* and *emphatic* correlate with the level measure L.

L and F are considered as competing measures for level, and M-L, M-F, H-L and H-F are considered as competing linguistic measures for span. As can be seen in table 5.4 by the coefficients marked with a tick to indicate the measure showing the strongest correlation with each respective feature, the L feature seems to be the most successful measure of level. This is a slightly different result compared to the result in experiment 2, but the results for the current experiment as well as for experiment 2 are very close and there is not much difference in either measure of level. Results for span show that

| | Level | | | | Span | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Feature | meanf0 | | medianf0 | | ± 2sds mean | | 90% Range | | 80% Range | |
| confident | 0.483 | | 0.483 | | **0.652** | | **0.691** | √ | **0.652** | |
| tense | -0.394 | | -0.394 | | **−0.845** | √ | **−0.732** | | **−0.845** | √ |
| harsh | 0.000 | | -0.154 | | -0.617 | | -0.309 | | **−0.694** | √ |
| expressive | 0.551 | | 0.551 | | **0.776** | | **0.801** | √ | **0.751** | |
| deep | **−0.868** | = | **−0.868** | = | -0.447 | | **−0.753** | √ | -0.345 | |
| weak | -0.567 | | **−0.693** | √ | -0.567 | | **−0.693** | √ | -0.504 | |
| irritated | **−0.630** | = | **−0.630** | = | **−0.756** | = | **−0.756** | = | **−0.630** | |
| happy | 0.327 | | 0.436 | | **0.846** | | **0.736** | | **0.873** | √ |
| afraid | **−0.732** | = | **−0.732** | = | -0.507 | | -0.620 | | -0.394 | |
| relaxed | **0.730** | √ | **0.626** | | **0.769** | | **0.861** | √ | **0.626** | |
| emphatic | **0.724** | √ | **0.651** | | **0.761** | | **0.798** | √ | **0.651** | |
| bored | **−0.655** | = | **−0.655** | = | **−0.718** | | **−0.794** | √ | -0.580 | |

**Table 5.5.** Results of correlation analyses for 2 long term distributional measures of level (measured in Hz) and 4 long term distributional measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the Railways passage. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

it is very difficult to establish which measure is the most successful in this replication experiment. Again the M-L measure is successful, but the success is matched, and maybe even slightly surpassed by the M-F measure.

Table 5.5 shows the correlation coefficient results comparing the Railways passage mode results with the long term distributional measures of level, measured in Hertz and the long term distributional measures of span, measured in semitones. Again, just these measures are featured in this study due to the results established in experiment 2. All $\rho$ that are significant ($p < 0.05$) are in a bold font. For example, table 5.5 shows that the features *deep, irritated, afraid, relaxed, emphatic* and *bored* correlate with the level measure mean f0.

Mean f0 and median f0 are considered as competing measures for level, and ± 2sds mean, 90% Range and 80% Range are considered as competing measures for span. As can be seen in table 5.5 by the coefficients marked with a tick to indicate the measure

showing the strongest correlation with each respective feature that the mean f0 feature seems to be the most successful measure of level. Results for span show that the 90% Range measure is the most successful, in complete contrast with the results of experiment 2 where ± 2sds mean was clearly the most successful. This suggests a lack of consistency in the ability of the long term distributional measures at capturing range effects.

| | Level (Hertz) | | | Span (Semitones) | | | |
|---|---|---|---|---|---|---|---|
| Feature | meanf0 | medianf0 | F | ± 2sds mean | 90% Range | 80% Range | M-L |
| confident | 0.483 | 0.483 | 0.274 | **0.652** | **0.691** | **0.652** | **0.717** √ |
| tense | -0.394 | -0.394 | -0.056 | **−0.851** √ | **−0.732** | **−0.845** | **−0.845** |
| harsh | 0.000 | -0.154 | 0.309 | -0.617 | -0.309 | **−0.694** √ | -0.463 |
| expressive | 0.551 | 0.551 | 0.350 | **0.776** | **0.801** | **0.751** | **0.901** √ |
| deep | **−0.868** = | **−0.868** = | **−0.702** | -0.447 | **−0.753** √ | -0.345 | -0.447 |
| weak | -0.567 | **−0.693** √ | -0.567 | -0.567 | **−0.693** | -0.504 | **−0.756** √ |
| irritated | **−0.630** = | **−0.630** = | -0.252 | **−0.756** = | **−0.756** = | **−0.630** | **−0.630** |
| happy | 0.327 | 0.436 | 0.027 | **0.846** | **0.736** | **0.873** = | **0.873** = |
| afraid | **−0.732** = | **−0.732** = | **−0.732** = | -0.507 | **−0.620** = | -0.394 | **−0.620** = |
| relaxed | **0.730** √ | **0.626** | 0.443 | **0.769** | **0.861** √ | **0.626** | **0.809** |
| emphatic | **0.724** √ | **0.651** | 0.491 | **0.761** | **0.798** = | **0.651** | **0.798** = |
| bored | **−0.655** = | **−0.655** = | -0.239 | **−0.718** | **−0.794** √ | -0.580 | -0.580 |

**Table 5.6.** Comparison of the correlation coefficients between 2 level measures based on long term distributional properties and 1 level measure based on linguistic properties and a comparison of correlation coefficients between 3 span measures based on long term general distributional properties and 1 span measure based on linguistic properties. Results in this table are based on the Railways passage.

The results in table 5.6 complete the analysis of the small scale replication of experiment 2. The results in table 5.6 shows that again, all the measures for level seem to be capable of capturing level effects, and again the M-L is the best measure of span. Results show that the 90% Range measure could be claimed to do as well as the M-L measure in this experiment. Combining the results of experiment 2 and the current experiment indicates that only the M-L measure of span consistently captures variations in listener judges' evaluations of speaker characteristics.

| feature | H | M | AW | KW | KG | TM | SS | HM | JC | JS |
|---|---|---|---|---|---|---|---|---|---|---|
| confident | | | 5 | 2 | 2 | 5 | 3 | 2 | 2 | 3 |
| confident | + | | 5 | 3 | 2 | 5 | 5 | 3 | 2 | 5 |
| confident | | + | 5 | 6 | 3 | 5 | 2 | 3 | 2 | 5 |
| confident | + | + | 5 | 5 | 2 | 5 | 5 | 2 | 2 | 5 |
| tense | | | 4 | 5 | 5 | 2 | 5 | 6 | 5 | 1 |
| tense | + | | 5 | 3 | 6 | 1 | 3 | 1 | 3 | 3 |
| tense | | + | 3 | 3 | 6 | 2 | 4 | 3 | 7 | 1 |
| tense | + | + | 2 | 3 | 4 | 3 | 2 | 1 | 6 | 2 |
| harsh | | | 3 | 2 | 2 | 4 | 1 | 1 | 1 | 1 |
| harsh | + | | 5 | 2 | 2 | 3 | 1 | 1 | 1 | 3 |
| harsh | | + | 3 | 1 | 1 | 3 | 1 | 2 | 2 | 1 |
| harsh | + | + | 6 | 3 | 2 | 4 | 1 | 1 | 2 | 1 |
| expressive | | | 2 | 1 | 3 | 4 | 3 | 1 | 1 | 3 |
| expressive | + | | 3 | 5 | 3 | 3 | 5 | 1 | 1 | 5 |
| expressive | | + | 3 | 5 | 3 | 3 | 3 | 3 | 2 | 5 |
| expressive | + | + | 5 | 2 | 3 | 5 | 4 | 1 | 2 | 5 |
| deep | | | 2 | 6 | 2 | 6 | 2 | 1 | 3 | 7 |
| deep | + | | 3 | 5 | 2 | 4 | 3 | 5 | 1 | 5 |
| deep | | + | 2 | 6 | 2 | 5 | 1 | 3 | 2 | 6 |
| deep | + | + | 1 | 5 | 2 | 6 | 1 | 1 | 2 | 5 |
| weak | | | 3 | 2 | 4 | 1 | 6 | 7 | 5 | 1 |
| weak | + | | 1 | 2 | 3 | 1 | 4 | 3 | 5 | 2 |
| weak | | + | 2 | 2 | 3 | 2 | 3 | 3 | 6 | 1 |
| weak | + | + | 1 | 1 | 3 | 2 | 3 | 3 | 6 | 1 |
| irritated | | | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 1 |
| irritated | + | | 1 | 3 | 2 | 3 | 2 | 1 | 2 | 1 |
| irritated | | + | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| irritated | + | + | 1 | 1 | 1 | 3 | 1 | 1 | 3 | 2 |
| happy | | | 1 | 2 | 1 | 2 | 2 | 3 | 1 | 3 |
| happy | + | | 1 | 2 | 1 | 2 | 4 | 1 | 1 | 5 |
| happy | | + | 1 | 2 | 2 | 3 | 1 | 2 | 1 | 4 |
| happy | + | + | 1 | 4 | 3 | 3 | 3 | 1 | 2 | 4 |
| afraid | | | 2 | 1 | 3 | 1 | 3 | 3 | 3 | 1 |
| afraid | + | | 1 | 1 | 3 | 1 | 2 | 1 | 3 | 2 |
| afraid | | + | 2 | 1 | 1 | 2 | 1 | 2 | 6 | 1 |
| afraid | + | + | 1 | 1 | 5 | 1 | 1 | 1 | 6 | 2 |
| relaxed | | | 2 | 2 | 2 | 3 | 2 | 1 | 1 | 4 |
| relaxed | + | | 2 | 6 | 1 | 5 | 2 | 1 | 3 | 4 |
| relaxed | | + | 1 | 3 | 1 | 2 | 1 | 5 | 1 | 5 |
| relaxed | + | + | 3 | 3 | 2 | 5 | 5 | 5 | 1 | 4 |
| emphatic | | | 4 | 3 | 3 | 1 | 2 | 2 | 1 | 3 |
| emphatic | + | | 1 | 4 | 2 | 3 | 4 | 3 | 1 | 5 |
| emphatic | | + | 4 | 5 | 3 | 5 | 2 | 2 | 2 | 5 |
| emphatic | + | + | 4 | 3 | 3 | 5 | 3 | 1 | 1 | 4 |
| bored | | | 5 | 2 | 1 | 3 | 2 | 3 | 6 | 4 |
| bored | + | | 2 | 2 | 4 | 5 | 2 | 4 | 7 | 2 |
| bored | | + | 3 | 4 | 2 | 2 | 3 | 2 | 2 | 2 |
| bored | + | + | 1 | 4 | 2 | 3 | 2 | 7 | 2 | 1 |

**Table 5.7.** mode for speakers in four conditions

| feature | H | M | AW | KW | KG | TM | SS | HM | JC | JS |
|---|---|---|---|---|---|---|---|---|---|---|
| confident | | | 5 | 2 | 2 | 5 | 3 | 2 | 2 | 3 |
| confident | + | | 5 | 3 | 2 | 5 | 5 | 3 | 2 | 5 |
| confident | | + | 5 | 6 | 3 | 5 | 2 | 3 | 2 | 5 |
| confident | + | + | 5 | 5 | 2 | 5 | 5 | 2 | 2 | 5 |
| tense | | | 4 | 5 | 5 | 2 | 5 | 6 | 5 | 1 |
| tense | + | | 5 | 3 | 6 | 1 | 3 | 1 | 3 | 3 |
| tense | | + | 3 | 3 | 6 | 2 | 4 | 3 | 7 | 1 |
| tense | + | + | 2 | 3 | 4 | 3 | 2 | 1 | 6 | 2 |
| harsh | | | 3 | 2 | 2 | 4 | 1 | 1 | 1 | 1 |
| harsh | + | | 5 | 2 | 2 | 3 | 1 | 1 | 1 | 3 |
| harsh | | + | 3 | 1 | 1 | 3 | 1 | 2 | 2 | 1 |
| harsh | + | + | 6 | 3 | 2 | 4 | 1 | 1 | 2 | 1 |
| expressive | | | 2 | 1 | 3 | 4 | 3 | 1 | 1 | 3 |
| expressive | + | | 3 | 5 | 3 | 3 | 5 | 1 | 1 | 5 |
| expressive | | + | 3 | 5 | 3 | 3 | 3 | 3 | 2 | 5 |
| expressive | + | + | 5 | 2 | 3 | 5 | 4 | 1 | 2 | 5 |
| deep | | | 2 | 6 | 2 | 6 | 2 | 1 | 3 | 7 |
| deep | + | | 3 | 5 | 2 | 4 | 3 | 5 | 1 | 5 |
| deep | | + | 2 | 6 | 2 | 5 | 1 | 3 | 2 | 6 |
| deep | + | + | 1 | 5 | 2 | 6 | 1 | 1 | 2 | 5 |
| weak | | | 3 | 2 | 4 | 1 | 6 | 7 | 5 | 1 |
| weak | + | | 1 | 2 | 3 | 1 | 4 | 3 | 5 | 2 |
| weak | | + | 2 | 2 | 3 | 2 | 3 | 3 | 6 | 1 |
| weak | + | + | 1 | 1 | 3 | 2 | 3 | 3 | 6 | 1 |
| irritated | | | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 1 |
| irritated | + | | 1 | 3 | 2 | 3 | 2 | 1 | 2 | 1 |
| irritated | | + | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| irritated | + | + | 1 | 1 | 1 | 3 | 1 | 1 | 3 | 2 |
| happy | | | 1 | 2 | 1 | 2 | 2 | 3 | 1 | 3 |
| happy | + | | 1 | 2 | 1 | 2 | 4 | 1 | 1 | 5 |
| happy | | + | 1 | 2 | 2 | 3 | 1 | 2 | 1 | 4 |
| happy | + | + | 1 | 4 | 3 | 3 | 3 | 1 | 2 | 4 |
| afraid | | | 2 | 1 | 3 | 1 | 3 | 3 | 3 | 1 |
| afraid | + | | 1 | 1 | 3 | 1 | 2 | 1 | 3 | 2 |
| afraid | | + | 2 | 1 | 1 | 2 | 1 | 2 | 6 | 1 |
| afraid | + | + | 1 | 1 | 5 | 1 | 1 | 1 | 6 | 2 |
| relaxed | | | 2 | 2 | 2 | 3 | 2 | 1 | 1 | 4 |
| relaxed | + | | 2 | 6 | 1 | 5 | 2 | 1 | 3 | 4 |
| relaxed | | + | 1 | 3 | 1 | 2 | 1 | 5 | 1 | 5 |
| relaxed | + | + | 3 | 3 | 2 | 5 | 5 | 5 | 1 | 4 |
| emphatic | | | 4 | 3 | 3 | 1 | 2 | 2 | 1 | 3 |
| emphatic | + | | 1 | 4 | 2 | 3 | 4 | 3 | 1 | 5 |
| emphatic | | + | 4 | 5 | 3 | 5 | 2 | 2 | 2 | 5 |
| emphatic | + | + | 4 | 3 | 3 | 5 | 3 | 1 | 1 | 4 |
| bored | | | 5 | 2 | 1 | 3 | 2 | 3 | 6 | 4 |
| bored | + | | 2 | 2 | 4 | 5 | 2 | 4 | 7 | 2 |
| bored | | + | 3 | 4 | 2 | 2 | 3 | 2 | 2 | 2 |
| bored | + | + | 1 | 4 | 2 | 3 | 2 | 7 | 2 | 1 |

Table 5.3: Ratings by listeners for speakers in four conditions

| Feature | All | AW | KW | KG | TM | SS | HM | JC | JS |
|---|---|---|---|---|---|---|---|---|---|
| confident | -0.823 | **−2.132** | -0.566 | -0.061 | -0.730 | **−2.203** | *−1.634* | -1.081 | *−1.715* |
| tense | *−1.902* | -1.354 | -0.476 | -1.286 | -0.119 | -1.199 | *−1.799* | -0.791 | 0.000 |
| harsh | -0.517 | -1.253 | -0.543 | -1.594 | -0.183 | -0.966 | *−1.623* | -1.063 | -0.368 |
| expressive | -0.871 | -0.604 | -0.423 | -0.431 | *−1.715* | -1.558 | **−1.983** | -0.108 | -0.144 |
| deep | -0.092 | -0.857 | -0.811 | -0.172 | -0.940 | -0.957 | -0.180 | 0.000 | -1.497 |
| weak | -1.460 | -2.354 | -0.531 | **−1.983** | -0.378 | -2.388 | -0.987 | -0.424 | -1.518 |
| irritated | -1.500 | 0.000 | -0.412 | -0.276 | -0.313 | -0.108 | -0.851 | -0.514 | -1.000 |
| happy | -1.353 | -0.566 | *−1.867* | -1.140 | -0.551 | -1.318 | *−1.930* | -0.359 | -0.520 |
| afraid | -0.707 | **−2.428** | -0.431 | -1.513 | -1.511 | -1.273 | -0.796 | -0.838 | *−1.890* |
| relaxed | -0.535 | -1.087 | -0.637 | -1.382 | -0.954 | -1.194 | -0.600 | -1.000 | -0.657 |
| emphatic | -1.216 | -1.187 | -1.018 | -0.351 | *−1.843* | *−1.813* | -1.382 | -1.387 | -0.551 |
| bored | -0.376 | -0.776 | -0.979 | -0.638 | -0.705 | -0.052 | -0.679 | -0.704 | -0.905 |

**Table 5.8.** Results of the Wilcoxon matched pairs test comparing listeners' responses for 12 adjectives with resynthesised speech for 8 speakers, between span increased at only non-sentence-intial peaks with span increased at both sentence-intial peaks and non-sentence-initial peaks. In this table, all Z coefficients that reach at least a significance level of $p < 0.05$ are in bold. All Z coefficients that reach a significance level between $p < 0.1$ and $p < 0.05$ are in italics.

### 5.4.2 Results of the Resynthesis Study

The *mode* for each feature for each of the four versions of each speaker's resynthesised voice averaging across all listeners were calculated using the SPSS statistical package. The full results can be found in table 5.7. In table 5.7 results show that for the feature *confident* speaker AW, TM and JC were rated as sounding the same no matter what span manipulations were made. Speaker KW was rated as sounding more confident as the pitch span was made wider, which would be as predicted given the assumption that increased span would be expected to lead to more positive judgements by listeners.

Of particular interest in the current experiment is whether there are any patterns which would show that a more complex model of pitch range, which would include both M and H as integral to the characterisation of span, would be of benefit in describing the communication of a selection of speaker characteristics. Table 5.8 shows

Z coefficients resulting from the Wilcoxon matched pairs test. All listener judges' responses for the "increased M" speech versions were compared to the responses for the "increased M and increased H" speech versions. Firstly the listener judges' responses were compared altogether, and then for individual speakers. None of the coefficients listed in the "All" column in table 5.8 reached the required level of significance ($p < 0.05$). At best it could be said that the feature tense at least showed a trend that increasing H as well as M makes a speaker sound less tense ($p < 0.1$), but this is not a significant result. Looking at the results of the Wilcoxon matched pairs test in table 5.8 for individual speakers it can be seen by those numbers in a bold font (eg speaker AW, Z = -2.132, $p < 0.05$) that there are a few significant results, and the Z coefficients printed in italics indicate that there are some trends which are not significant. It is clear from these results that generally speaking, there is no regular pattern that would suggest that there is any need to have a more complex description of pitch span. Speakers are generally not rated any differently whether or not their span is increased at just the M target points or whether span is increased at both M and H target points.

Table 5.9 shows the results of further Wilcoxon matched pairs analyses. The responses of listener judges to version 1 of speakers' speech, which involved no increases in span, were compared to the responses of listener judges to version 4 of speakers' speech, which involved the increase of span at both the H and the M points on a speaker's contour. Listener judges' responses have been used to establish which measure of span most effectively captures differences across speakers in experiment 1 and 2 in this thesis. It is interesting to see whether variation in span within speakers would make predictable changes in listener judges' responses to certain characteristics. For this whole thesis it has been found that wider spans lead to more positive characterisations of speakers while narrower spans lead to more negative characterisations of speakers. While this finding has only been used in relation to cross-speaker differences, it would seem that such an assumption should also relate to within speaker differences. Results in the "All" column of table 5.9 shows that across all speakers

| Feature | All | AW | KW | KG | TM | SS | HM | JC | JS |
|---|---|---|---|---|---|---|---|---|---|
| confident | **−3.000** | -1.559 | -1.141 | -0.960 | -0.586 | **−2.013** | **−1.983** | -0.378 | -0.846 |
| tense | *−1.813* | -1.438 | -0.418 | -0.155 | -0.471 | **−2.368** | **−2.379** | **−2.121** | -0.312 |
| harsh | -1.61 | -1.372 | -0.795 | -0.106 | -0.689 | 0.000 | **−2.264** | *−1.724* | -1.421 |
| expressive | **−2.127** | -0.845 | -0.358 | -1.000 | -0.426 | -0.780 | -0.712 | -1.552 | *-1.845* |
| deep | *−1.833* | -0.351 | -1.496 | -0.690 | *−1.638* | -0.690 | -0.123 | -1.387 | -1.207 |
| weak | *−1.876* | **−2.263** | -0.284 | -0.212 | -0.302 | -1.551 | **−2.320** | -1.257 | -0.085 |
| irritated | -0.130 | -0.070 | 0.000 | -0.359 | -0.060 | -0.604 | *−1.628* | *−1.916* | -0.061 |
| happy | **−3.05** | -1.450 | -0.966 | -1.590 | -1.100 | -0.604 | -1.121 | -1.403 | **−2.401** |
| afraid | -1.231 | **−2.585** | -0.351 | -1.282 | *−1.706* | **−2.066** | **−2.375** | -1.508 | -0.816 |
| relaxed | -1.102 | -0.957 | -0.616 | 0.000 | -0.060 | -1.373 | **−2.251** | -1.403 | -0.354 |
| emphatic | *−1.752* | **−2.393** | -0.639 | -1.000 | -1.491 | -0.343 | -0.306 | -1.040 | -1.441 |
| bored | -1.361 | -0.359 | -0.564 | -0.778 | -1.252 | -0.690 | -0.205 | -1.543 | *−1.739* |

**Table 5.9.** Results of the Wilcoxon matched pairs test comparing listener's responses for 12 adjectives with resynthesised speech for 8 speakers, between no span increase with span increased at both sentence-intial peaks and non-sentence-initial peaks. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. All Z coefficients that reach a significance level between $p < 0.1$ and $p < 0.05$ are in italics.

as a whole the wider spans in version 4 were rated as being significantly more confident, expressive and happy than the narrower version 1 speech. Also there are trends (though not significant) that the wider spans in version 4 were rated as being more emphatic than the version 1 speech and that the wider spans were rated as less tense and weak. Results for individual speakers are also reported. The fact that results are significant for some speakers and not others, just indicates that pitch range is only one of many variables with which speakers can be characterised. The results in table 5.9 can be interpreted as showing that pitch range isn't as clear a variable as others for some speakers in how they are characterised by listeners.

## 5.5 Conclusions and Discussion

Experiment 3 is divided up into 2 separate sections. Firstly a small scale replication of experiment 2 was carried out. In this "replication" study, there was only one passage

and one type of speech presentation, but again there was an equal number of Scots and English, men and women, 8 speakers in all, as opposed to the 32 speakers used in experiment 2. In the "replication" study scale of measurement was not investigated. It was assumed from the results of experiment 2 that level should be measured in Hertz and span should be measured using musical semitones. Results for level again suggested that there is no clear advantage in using any one particular measure of level. F, L and the long term distributional properties all capture the differences in speaker level effectively.

Results for span show that for the competing linguistic measures of span, M-L does not stand out on its own as the most successful measure of span; the M-F measure also accounts for span successfully. Comparing the linguistic measures with the long term distributional measures shows that for this small scale experiment, the measure based on the 90% range of f0 successfully accounts for the difference in listener judges' perception of speaker characteristics. Rather than putting a hole in the argument that the measure of span should be based on tonal targets in speech, results for the two experiments show that the M-L measure shows consistency in accounting for speaker variation, whereas the long term distributional measures show no clear patterns in success of accounting for speaker variation. In experiment 2 the range measure based on 4 standard deviations around the mean was the only measure close to being successful, and the 90% range and 80% range measures were not at all satisfactory. In experiment 3 the 90% range measure is the most successful. The results of the two experiments together show that only one type of measure can consistently be relied upon to measure speaker span variation. Adding the cross-speaker findings of this thesis to the results of within-speaker range variation found in Shriberg *et al.* (1996), there is now a clear starting point for how to measure pitch range for future studies interested in this pitch variable.

Results of the "resynthesis" study reported in the current chapter show that while it is important to understand how to measure pitch range and to describe what pitch range actually is, it is important not to overestimate its importance. The results of

the resynthesis experiment confirm the original hypothesis that pitch range is a global speaker characteristic, and fine-grained manipulation of the parameters discussed in this thesis do not seem to capture any more fine-grained details in capturing differences in speaker characteristics. It can still be said that wide pitch spans are characterised more postitively than narrower spans, but results of the resynthesis experiment suggests that a model of finer detail will not be able to tease apart more specific characteristics that will seperate a speaker from being perceived as, for example, happy as opposed to emphatic.

# Chapter 6

# Conclusions

## 6.1 Thesis Review

This thesis has aimed to address the lack of coherence in research that involves a notion of pitch range. From the outset, three strands of research - extralinguistic, paralinguistic and linguistic - were defined in section 1.2. Within these three areas of research, pitch range has been used to explain, or partly explain, certain phenomena. Examples of research that have used a notion of pitch range, as discussed in chapter 1, include the lowering of mean pitch in male speech between the ages of 20 through to 40, then a rise in mean pitch from age 60 upwards (Hollien & Shipp 1972) through to the Beckman & Pierrehumbert (1986) reanalysis to H*..H* of the original Pierrehumbert (1980) H*+H..H* tonal sequence (cf. section 1.4.2).

The key problem area identified in chapters 1 and 2 is that pitch range has been afforded multiple definitions both within and across research disciplines. In this thesis pitch range has been characterised in terms of two independent measures, span and level, following Ladd (1996). The majority of the definitions used have been based on long term distributional properties. The benefits of long term distributional measures are that they are simple measures to extract data for. This though is not a principled

reason to accept them as suitable measures. The long term distributional measures are oversimplified in the sense that they do not capture variations in pitch range effectively. Huttar (1968) first showed that pitch range expansion occurs from the bottom of the range upwards, but the long term distributional properties use the middle of the range as the starting point and expand both up and down the range (Brown *et al.* 1973).

## 6.2 Overview of Results

Of all the studies involving pitch range, across the 3 research areas - extralinguistic, paralinguistic and linguistic - the recent studies by Ladd & Terken (1995) and Shriberg *et al.* (1996) identify a potential measure of range that could unify an account of pitch range. Measures of level and span were based on specific target points in speech that are linguistic in nature. Shriberg *et al.* (1996) showed that variation in these targets were predictable, within-speaker, from a normal condition to a "speaking up" condition, using their model.

The Shriberg *et al.* (1996) model was the point of departure for the experiments reported in this thesis. Firstly the 2 parameter model based on Dutch speech was used to test whether the characterisation of pitch range was also able to capture across-speaker differences. Results from chapter 3 show that the variation of level and span based on tonal targets in speech captures variations in listener judgements of speaker characteristics. The results of chapter 3 show that a more refined approach to pitch range also shows similar patterns to previously established results in this area of research (Huttar 1968, Mozziconacci 1998).

Having established a method of measuring span and level that could contribute to voice and linguistic research, the measures of level and span based on Dutch speech were extended to English speech, to investigate whether similar success could be had

in establishing a model of pitch range that would be suitable for all needs. The extended study, using English speech, as reported in chapter 4, was able to identify which tonal targets in speech best characterise pitch range, given the chosen methodology. Two comparative studies were carried out: first a comparison of the linguistic measures of span and level with the measures based on the long term distributional properties of f0; second a comparison of three possible scales of measurement, linear Hertz, logarithmic musical semitones and the ERB-rate scale based on the frequency selectivity of the auditory system. Results show that the linguistic model is better at characterising range across speakers than the competitor long term distributional models. Specifically, the "best" topline is the average of a speaker's non-sentence-intial highs, the best bottomline is the average of a speaker's post-accent valleys, and level is best characterised by sentence final low (cf. figure 4.1). The musical semitone scale is the best unit of measure for span, and there is no difference between Hertz and ERB for measuring level (for which semitones are not a suitable scale due to its logarithmic nature). These results were partially replicated in a small scale experiment reported in chapter 5. Although one of the long term distributional measures matched the linguistic measure in the replication, the only consistently successful measure has been based on tonal targets in speech.

A third experiment, also reported on in chapter 5, was carried out to see how the linguistic measures of pitch range, developed by their success at capturing the variation in the perception of speaker characteristics across-speakers, using English speech, could be used to influence the perception of speaker characteristics within-speakers. In manipulating the pitch span parameters using resynthesis, it was shown that there is a general tendency for pitch span increase to lead to more positive ratings of speakers. This result is entirely consistent with previous results (Uldall 1960, Brown *et al.* 1974). Furthermore, the third experiment was able to test whether a more detailed model of pitch range would be able to capture more specific variations in the perception of speaker characteristics. Results showed that a more detailed description

of pitch span generally had no impact in teasing apart the finer details of the perception of speaker characteristics. Results showed that while pitch range has an effect on speaker characterisation along a positive-negative continuum, it is only partly responsible for the variation in speaker characteristics, which may also be encoded within variations of speech quality, intensity and duration.

## 6.3 Discussion

Given the results from the 3 experiments reported on in this thesis, are we any closer to having a unified account of pitch range with a sound theoretical basis, and with application to studies in voice and linguistics research?

### 6.3.1 Voice

Shriberg *et al.* (1996) propose a model of pitch range which relates the topline and the bottomline of span and sets a level to tonal targets in speech. Using these target points, they showed that the manipulation of these targets, when asking a speaker to "speak-up", are predictable within speaker. The very predictability of these targets is strong evidence that they can be related to a notion of pitch range, which has been lacking an explicit definition. Results from the experiments reported in this thesis show that a pitch range model based on tonal targets in speech can also account for cross-speaker differences in pitch range. An important contribution of the model of pitch range established in this thesis is that it not only captures differences in pitch range, but also provides further evidence for how pitch range is manipulated for communicating affect. It captures the established pattern of pitch range in which range expands from the bottom up (Huttar 1968, Shriberg *et al.* 1996). The pattern in pitch range expansion is not clearly captured in any model of pitch range based on long term distributional properties of f0. In fact, there is no principled motivation offered for any of the long

term distributional measures used in this thesis; these long term distributional measures being only used for demonstration purposes. While it is clear that all long term distributional measures will capture a topline somewhere near the top of a speaker's f0 distribution and bottomline somewhere near the bottom of a speaker's f0 distribution, the motivation for selection seems more to depend on which measure best eradicates f0 data that are just outliers due to octave errors. Machine error is not the most positive of motivations for a measure that should have a theoretical basis.

One of the key motivations for this thesis has been to show that progress can be made in a discussion of pitch range by the integration of results from linguistic, paralinguistic and extralinguistic research. Essentially it is imperative to have a clear picture of the whole system of pitch and what affects it, to be able to fully understand important issues that may be restricted to one specific field. The model of pitch proposed by Liberman & Pierrehumbert (1984), in order to decide linguistic issues in the study of intonation, not only attempted to model issues of *tune* but also tried to incorporate the effects of *prominence, declination* and *pitch range* (discussed in section 1.4.1). One of the key aspects of the Liberman & Pierrehumbert (1984) model is the domain over which pitch range variation operates. Liberman & Pierrehumbert (1984) describe pitch range as minimally varying at a phrasal level, whether these variations are reflective of linguistic features such as the instantiation of various tones, or more paralinguistic in nature such as the raising of the voice in "speaking up". Tonal features are related to a reference level which is a speaker specific level set above the baseline with H accents modelled at some distance above the reference level and L accents at some distance below the reference level. The reference level itself can be moved up or down to capture paralinguistic effects such as the raising of the voice. One of the key aspects of the model is that the reference level is of primary importance with the additional constraint of a floor effect being the bottom of the range, which is constant (cf. section 1.4.2).

The Shriberg *et al.* (1996) model, on which the experimental work reported in this thesis is based, is similar in spirit to the model proposed by Liberman & Pierrehumbert

(1984). Again the motivation is to provide an integrated model of pitch that can account for linguistic, paralinguistic and extralinguistic features. The difference between the two models is that the Shriberg *et al.* (1996) model has a stronger bias to pitch range being made up of two components, *level* and *span*, and that *span* is characterised by a top-line AND a bottom-line, not just a single reference line. There is a simple way in distinguishing the priorities of the two models. The Liberman & Pierrehumbert (1984) version models pitch features on their all important reference line with a floor effect of the constant bottom of the range. The Shriberg *et al.* (1996) version models pitch features from the constant bottom of the range. Given the invariance of the bottom of the range, and the results suggesting that expansion of pitch range occurs from the bottom up, it seems clear that a model of pitch features should be projected from the stable bottom of the range position.

### 6.3.2 Linguistics

**The issue of pitch frameworks**

The results of the experiments reported in this thesis support the view that abstractions such as the Gårding (1983) tonal grid and references to toplines in pitch range should be constrained by the linguistic specifications of accent peaks, as suggested by Ladd & Cutler (1983).

Ladd (1996:252-257) discusses two conceptual frameworks in which to describe pitch features: the initialising approach, in which pitch features are described in relation to other parts of an utterance and the normalising approach, in which pitch features are described relative to the speaker's voice. While an initialising approach was used with some degree of success by Crystal (1969), such an approach suffers from clear theoretical problems, as highlighted by data taken from Connell & Ladd (1990). A key point is provided by data taken from Connell & Ladd (1990) in their study of Yoruba. Using the initialising approach, pitch features can only be described relative

to different pitch features in an utterance such that a high range pitch feature is "high" compared to a mid range pitch feature which is "mid" compared to a low range pitch feature. There is no possible characterisation of high range features if there is no different type of range feature to compare it to; a situation that occurs regularly in Yoruba. In Yoruba, utterances can be composed just using strings of the same tones throughout the utterance. Clearly the initialising approach cannot account for a string of same tone utterances as it is variation within an utterance that gives each tone its character.

A normalising approach "reifies the notion of 'pitch range' in terms of some speaker reference points such as upper and lower f0 values. Such a model attempts to abstract away from differences between speakers, paralinguistic effects and so on, and express the invariant characterisations of tones in terms of the idealised speaker range from this process of factoring out sources of variation." (Ladd 1996). This thesis has shown that "M" and "L" are suitable upper and lower values for capturing differences between speakers and have been shown to be linked with the perception of speaker characteristics.

Two general versions of the normalising approach have been suggested. In the following extended quote, Rose (1987) says "The two most common strategies in the normalisation and scaling of f0 are of the general linear form:

- $f0_{norm} = (f0_i - f0_{ref})/f0_{range}$

Fraction of range (FOR) transforms (eg Earle (1975), Takefuta (1975), Rose (1982), Ladd *et al.* (1985)) express an observed f0 value as a fraction of the difference between two range-defining f0 values, eg.

- $f0_{norm} = (f0_i - f0_{min})/(f0_{max} - f0_{min})$.

Z-score transforms (eg. Jassem (1975), Menn & Boyce (1982)), express an observed f0 value as a multiple of a measure of dispersion away from a mean f0 value, eg.

- $f0_{norm} = (f0_i - \overline{f0})/s,$

where $s$ is one standard deviation about the mean f0 ($\overline{f0}$)."

Both the Liberman & Pierrehumbert (1984) and the Shriberg *et al.* (1996) models are based on a normalising approach which characterise the level of pitch features relative to some speaker specific reference point, characteristic of the FOR method of normalisation. Clearly the Liberman & Pierrehumbert (1984) model tries to characterise pitch features around the all important speaker-specific "reference level" with a pitch floor imposed by the bottom of the range specification. The Shriberg *et al.* (1996) model takes the more traditional format of defining pitch features between two range defining features. This thesis proposes that the range-defining values for the FOR method of normalisation should be based on "M" and "L". Although questions have been raised against the FOR method of normalisation (Rose 1987) it has also been shown that long term mean and standard deviation are not effective as normalisation parameters (Rose 1991). Again results from this thesis suggest that long term distributional properties of f0 do not effectively characterise pitch range, and if it is possible to get a closer approximation to the linguistic range measure, possibly by including skew and kurtosis data, then the issue of normalisation parameters could be solved.

**The issue of the H*+H...H* reanalysis**

In section 1.4.1 a definition of pitch range as "a global, or at least phrase-sized choice of pitch scaling parameters" (Liberman & Pierrehumbert 1984) was introduced. This definition was mentioned in the discussion of the reanalysis of the H*+H...H* sequence introduced by Beckman & Pierrehumbert (1986). In the 1986 analysis of the problematic tonal sequence under discussion, the H*+H...H* analysis was replaced with just two H* accents and the sustained high transition was explained as an effect of a local elevation of pitch level and compression of pitch span, as described in figure 1.4.

Two key issues arise from the Beckman & Pierrehumbert (1986) reanalysis. The first

issue concerns the domain over which pitch range can be specified. There is a clear conflict between first describing pitch range first as a global measure, then using pitch range manipulation at a very local level to describe a problematic tonal feature. Experiment 3 has contributed to the issue of how large the domain in which pitch range variation can have an affect, by having a closer look at the influence of the extra high peaks that occur sentence-initially. Results from experiment 3 show that the sentence intial highs seem not to have any influence on the perception of speaker characteristics above and beyond the global top line set by averaging the non-initial peaks. These results are taken as further evidence to suggest the global nature of pitch range and pitch range manipulations, and consequently is taken as further evidence against the reanalysis of the H*+H...H* tonal sequence to H*...H* made by Beckman & Pierrehumbert (1986). So the reanalysis has to be seen as doubtful because pitch range is set for a more substantial domain than that required for the reanalysis to be acceptable. Even if the domain over which pitch range can vary is reduced to something more small scale, that domain should at least be definable and the isolated area that is encapsulated by the "+H" segment does not meet that requirement.

The second issue and more fundamental problem with the H*+H analysis is the notion of "raised compressed pitch range". There is no evidence to suggest that such a pitch range manipulation takes place. One of the clearest facts about speakers' voices is that the bottom of the range is far more stable than the top end of the range. The sentence final low is so stable that it is considered to be a speaker constant, but also low tones are more stable than high tones. Given the model defined in this thesis it is theoretically possible for range to be raised and compressed. This would rely on the measures of span and level to be totally independent, such that the "F" level would remain low, while the "L" bottomline of span was close the "M" topline of span in a speaker specific, elevated position. But evidence from Shriberg *et al.* (1996) shows that when pitch range is raised at the bottom of the range, such a raising is generally accompanied by a widening in span, rather than a "compression". Also results from this thesis show that there is a very strong correlation between the "F" and

the "L" measures and that level and span are only partially independent of each other. Liberman & Pierrehumbert (1984) avoid the issue of having to describe huge leaps of f0 valleys up to the H tone level creating the effect of a sustained high, by only making the bottom of the range important as a floor for all pitch features, and describing all other pitch features relative to one speaker specific reference line. Given the wealth of data that exists about variation in pitch movements at the bottom of the range whether for linguistic or paralinguistic purposes (including Liberman & Pierrehumbert 1984, Shriberg *et al.* 1996 and results from this thesis), there is no evidence that supports such a huge leap in the bottom of the speaker range and such a compression of range at such a local level.

Clearly this leaves a problem for the tonal analysis of English. Initially there was a tonal analysis H*+H... H* (Pierrehumbert 1980) that was unacceptable for various reasons beyond the scope of this thesis. The difference between the H*+H... H* and the H*... H* contours was resolved by putting the burden of explanation outside the tonal description of English and onto a model of pitch range by Beckman & Pierrehumbert (1986). The reanalysis has not so much helped to resolve this issue, but is more characteristic of "passing the buck". Due to the results of this thesis it is clear that the matter of this tricky tonal problem still remains unresolved though the resolution of this tonal issue is beyond the scope of this thesis.

## 6.4 Further Research

Ideas and suggestions for further research can have two motivations. The first motivation is to find ways of improving the current set of studies reported on in this thesis so as to confirm the conclusions that have been drawn from the current set of data. The second motivation is to draw on the results of the current experiment and make suggestions to take the research to a new level.

There are a few aspects about the experiments reported in this thesis that are problematic. Firstly the results from experiment 2 showed that there were differences in the ratings of speakers based on accent type and sex of speaker, and also by sex of listener. It may well be advantageous to replicate the study but having even more control over these variables, possibly by eliminating them rather than balancing them which was the method chosen for the main experiment of this thesis. This would require a great deal of time and effort though, if the prerequiste of running a large scale study remains, which was an integral part of experiment 2.

The results of experiment 3 are only based on a small scale study, and the span manipulations by resynthesis were only impressionistic. To increase the validity of these results, more detailed attention to span manipulations would be necessary, and the conclusions drawn by the results of the resynthesis experiment reported on in chapter 5 would be supported by a larger-scale version of the study.

One issue that is more general in nature relates to the specification of the L measurements. We clearly stated in section 2.1.1 that we would be working with the Bruce & Gårding (1978) model in which H tones and L tones correspond to local maxima and local minima. This could be considered a simplistic model to follow given the subsequent model proposed by Pierrehumbert (1980). One of the key differences between the two models is in the phonological status of valleys. While all valleys would be measured as L in this thesis, if we were to follow the Pierrehumbert (1980) model we would have to take into consideration the different phonological statuses of *sag*, *L-* and *L**.

In Pierrehumbert's original system a low tone associated with an accented syllable is written as L*, a low tone preceeding or following a pitch accent is written as L-, while a sagging transition describes a valley transition between two H* accents which is distinct from an L- between two H* accents. If there is more emphasis placed on the second pitch accent, the intervening valley is described as L-. When no extra emphasis is placed on the second H*, the intervening valley is described as sag.

While it has been argued that pitch range is expanded from the bottom up (Shriberg *et al.* 1996), pitch range variation may well affect the three types of valley in different ways. For example, an increase in range may well lower the average L* to a level closer to the sentence final low, but leave L- and sag unaltered (Liberman & Pierrehumbert 1984, Pierrehumbert & Beckman 1988). Likewise, the bottom of a pitch span measure may well be slightly different if this measure was taken to be the average L* as opposed to the average L-. Due to the framework being used in this thesis, close attention was not made to the phonological variants that represent valleys in f0 contours. Because of this some degree of caution may well have to be shown in the conclusions drawn. On review of the speech recordings used, the vast majority of valleys measured were L- targets. We acknowledge that there is the possibility that if we had decided to use the Pierrehumbert (1980) framework, the results might be slightly different, but we are confident that the overall conclusions would remain the same.

One of the key next steps to take in continuing research, as opposed to fine tuning the current research, relates to the long term distributional measures of pitch range. This thesis has shown that there is a strong basis for relating pitch range to tonal targets in speech. Given this basis, it is clear what any measure of pitch range based on long term distributional properties should be aiming to estimate. In making an estimate for a topline that will be similar to the average of the non-initial sentence highs and an estimate for a bottomline similar to the average post-accent valleys, it is considered that use of skew and kurtosis data will be an important start in improving the previous methods, which place too much value in the assumption that f0 production is normally distributed. It has been shown in table A.7 that the patterns of f0 around the mean are very much speaker specific. The variations in skew and kurtosis would offer greater insight into the patterns of f0 around the mean and could perhaps be utilised to make better estimations of span and level for speakers, as previously discussed in section 4.6.

Murray & Arnott (1993), in a review of Lieberman & Michaels (1962) highlight that Lieberman & Michaels's results suggest that "fundamental frequency was not wholly

responsible for the communication of emotion". Results from general speaker characteristics support Murray & Arnott's conclusion. There are clearly identifiable variables in the signaling of speaker affect (Ladd *et al.* 1985), pitch range being one. Given a clearer understanding of pitch range, it should be possible to clarify the importance of pitch range in the communication of emotion. Murray & Arnott (1993) also state that "the intelligibility of speech synthesizers is approaching that of human speech." It is also important to establish how much a clearer understanding of pitch range, provided by this thesis, could influence the progress of speech technologists in controlling for more subtle voice changes in synthesis.

Finally, a goal to work towards should be to establish the best normalising model which will be able to "abstract away from differences between speakers, paralinguistic effects and so on, and express the invariant characterisations of tones..." (Ladd 1996).

# Appendix A

# Pitch Range Data for Experiment 2

| | | | Span (Hz) | | | | Level (Hz) | |
|---|---|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | H-F | H-L | M-F | M-L | F | L |
| AP | male | Scottish | 122.40 | 107.17 | 71.95 | 56.72 | 89.39 | 104.62 |
| HM | male | Scottish | 96.52 | 79.18 | 57.79 | 40.45 | 96.11 | 113.45 |
| JS | male | Scottish | 110.74 | 99.99 | 69.84 | 59.09 | 82.94 | 93.69 |
| AB | male | Scottish | 105.14 | 89.10 | 70.11 | 54.07 | 88.33 | 104.37 |
| GF1 | male | Scottish | 103.87 | 91.17 | 59.65 | 46.95 | 85.18 | 97.88 |
| FH | male | Scottish | 94.72 | 83.45 | 54.79 | 43.52 | 84.06 | 95.33 |
| TM | male | Scottish | 73.71 | 61.49 | 47.80 | 35.58 | 74.94 | 87.16 |
| KW | male | Scottish | 95.83 | 71.25 | 76.85 | 52.27 | 64.67 | 89.25 |
| SM | male | English | 101.86 | 86.40 | 52.12 | 36.66 | 83.72 | 99.18 |
| GF2 | male | English | 41.00 | 36.80 | 25.37 | 21.17 | 97.00 | 101.20 |
| RC | male | English | 95.42 | 86.78 | 46.82 | 38.18 | 84.83 | 93.47 |
| ME | male | English | 105.54 | 96.00 | 56.21 | 46.67 | 85.11 | 94.65 |
| RL | male | English | 79.82 | 69.58 | 48.07 | 37.83 | 85.29 | 95.53 |
| VR | male | English | 102.67 | 83.11 | 75.30 | 55.74 | 60.68 | 80.24 |
| JB | male | English | 130.47 | 103.14 | 99.84 | 72.51 | 79.94 | 107.27 |
| GB | male | English | 122.97 | 98.01 | 86.18 | 61.22 | 73.33 | 98.29 |
| FL | female | English | 77.86 | 58.89 | 48.17 | 29.20 | 147.61 | 166.58 |
| NG | female | English | 110.83 | 97.91 | 57.60 | 44.68 | 173.50 | 186.42 |
| SO | female | English | 222.51 | 195.96 | 97.95 | 73.40 | 155.33 | 179.88 |
| NC | female | English | 199.95 | 176.02 | 113.61 | 89.68 | 126.58 | 150.51 |
| JK | female | English | 151.47 | 117.68 | 100.89 | 67.10 | 142.79 | 176.58 |
| JV | female | English | 88.16 | 63.08 | 66.64 | 35.56 | 148.67 | 173.75 |
| RS | female | English | 127.03 | 111.05 | 77.67 | 61.69 | 124.22 | 140.20 |
| MT | female | English | 201.27 | 180.92 | 105.95 | 85.60 | 115.94 | 136.29 |
| JT | female | Scottish | 102.83 | 92.12 | 62.00 | 51.29 | 190.78 | 201.49 |
| JC | female | Scottish | 85.11 | 57.33 | 57.94 | 30.16 | 163.67 | 191.45 |
| KG | female | Scottish | 142.15 | 130.56 | 88.88 | 77.29 | 150.67 | 162.26 |
| DN | female | Scottish | 92.00 | 69.42 | 67.90 | 45.32 | 149.89 | 172.47 |
| SS | female | Scottish | 213.89 | 171.45 | 130.24 | 87.80 | 131.50 | 173.94 |
| AW | female | Scottish | 212.26 | 190.89 | 131.15 | 109.78 | 150.42 | 171.79 |
| JD | female | Scottish | 137.76 | 122.18 | 71.63 | 56.05 | 150.50 | 166.08 |
| JO | female | Scottish | 150.63 | 136.05 | 92.81 | 78.23 | 135.00 | 149.58 |

**Table A.1.** Span and level measures for each speaker from the English Speech Database, measured in Hz

| | | | Span (ERB) | | | | Level (ERB) | |
|---|---|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | H-F | H-L | M-F | M-L | L | F |
| AP | male | Scottish | 2.85 | 2.42 | 1.80 | 1.38 | 3.55 | 3.13 |
| HM | male | Scottish | 2.28 | 1.81 | 1.45 | 0.98 | 3.79 | 3.32 |
| JS | male | Scottish | 2.67 | 2.37 | 1.80 | 1.49 | 3.26 | 2.95 |
| AB | male | Scottish | 2.51 | 2.07 | 1.77 | 1.32 | 3.55 | 3.10 |
| GF1 | male | Scottish | 2.52 | 2.16 | 1.55 | 1.19 | 3.37 | 3.01 |
| FH | male | Scottish | 2.33 | 2.01 | 1.44 | 1.12 | 3.30 | 2.98 |
| TM | male | Scottish | 1.94 | 1.58 | 1.32 | 0.96 | 3.07 | 2.71 |
| KW | male | Scottish | 2.53 | 1.79 | 2.09 | 1.35 | 3.13 | 2.39 |
| SM | male | English | 2.49 | 2.05 | 1.38 | 0.94 | 3.41 | 2.97 |
| GF2 | male | English | 1.05 | 0.94 | 0.67 | 0.55 | 3.46 | 3.35 |
| RC | male | English | 2.34 | 2.10 | 1.24 | 1.00 | 3.25 | 3.00 |
| ME | male | English | 2.55 | 2.28 | 1.47 | 1.20 | 3.28 | 3.01 |
| RL | male | English | 2.00 | 1.71 | 1.27 | 0.98 | 3.31 | 3.02 |
| VR | male | English | 2.72 | 2.11 | 2.09 | 1.48 | 2.87 | 2.27 |
| JB | male | English | 3.09 | 2.33 | 2.48 | 1.71 | 3.63 | 2.86 |
| GB | male | English | 3.01 | 2.29 | 2.24 | 1.51 | 3.38 | 2.66 |
| FL | female | English | 1.61 | 1.18 | 1.04 | 0.61 | 5.05 | 4.63 |
| NG | female | English | 2.05 | 1.78 | 1.14 | 0.81 | 5.47 | 5.20 |
| SO | female | English | 3.80 | 3.26 | 1.93 | 1.40 | 5.34 | 4.80 |
| NC | female | English | 3.78 | 3.21 | 2.38 | 1.81 | 4.69 | 4.12 |
| JK | female | English | 2.90 | 2.14 | 2.05 | 1.30 | 5.27 | 4.51 |
| JV | female | English | 1.79 | 1.24 | 1.28 | 0.72 | 5.21 | 4.65 |
| RS | female | English | 2.64 | 2.25 | 1.72 | 1.33 | 4.45 | 4.06 |
| MT | female | English | 3.91 | 3.41 | 2.32 | 1.81 | 4.36 | 3.85 |
| JT | female | Scottish | 1.84 | 1.62 | 1.16 | 0.95 | 5.78 | 5.56 |
| JC | female | Scottish | 1.67 | 1.08 | 1.18 | 0.59 | 5.58 | 4.99 |
| KG | female | Scottish | 2.69 | 2.43 | 1.80 | 1.54 | 4.96 | 4.70 |
| DN | female | Scottish | 1.86 | 1.36 | 1.41 | 0.91 | 5.18 | 4.68 |
| SS | female | Scottish | 3.94 | 2.97 | 2.64 | 1.67 | 5.21 | 4.24 |
| AW | female | Scottish | 3.73 | 3.25 | 2.52 | 2.04 | 5.17 | 4.69 |
| JD | female | Scottish | 2.62 | 2.28 | 1.48 | 1.13 | 5.04 | 4.69 |
| JO | female | Scottish | 2.95 | 2.60 | 1.95 | 1.61 | 4.67 | 4.33 |

**Table A.2.** Span and level measures for each speaker from the English Speech Database, measured in ERB

| | | | Span (semitones) | | | |
|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | H-F | H-L | M-F | M-L |
| AP | male | Scottish | 14.93 | 12.21 | 10.22 | 7.50 |
| HM | male | Scottish | 12.04 | 9.17 | 8.15 | 5.28 |
| JS | male | Scottish | 14.68 | 12.57 | 10.58 | 8.47 |
| AB | male | Scottish | 13.57 | 10.68 | 10.12 | 7.23 |
| GF1 | male | Scottish | 13.80 | 11.40 | 9.19 | 6.78 |
| FH | male | Scottish | 13.06 | 10.89 | 8.69 | 6.51 |
| TM | male | Scottish | 11.86 | 9.24 | 8.54 | 5.93 |
| KW | male | Scottish | 15.74 | 10.16 | 13.56 | 7.98 |
| SM | male | English | 13.78 | 10.85 | 8.38 | 5.45 |
| GF2 | male | English | 6.10 | 5.37 | 4.02 | 3.29 |
| RC | male | English | 13.05 | 11.37 | 7.61 | 5.93 |
| ME | male | English | 13.96 | 12.12 | 8.78 | 6.94 |
| RL | male | English | 11.44 | 9.47 | 7.74 | 5.78 |
| VR | male | English | 17.14 | 12.31 | 13.97 | 9.13 |
| JB | male | English | 16.75 | 11.66 | 14.03 | 8.94 |
| GB | male | English | 17.05 | 11.98 | 13.45 | 8.38 |
| FL | female | English | 7.33 | 5.24 | 4.89 | 2.80 |
| NG | female | English | 8.55 | 7.31 | 4.96 | 3.72 |
| SO | female | English | 15.30 | 12.76 | 8.49 | 5.92 |
| NC | female | English | 16.41 | 13.41 | 11.09 | 8.09 |
| JK | female | English | 12.52 | 8.84 | 9.25 | 5.58 |
| JV | female | English | 8.06 | 5.36 | 5.92 | 3.22 |
| RS | female | English | 12.19 | 10.10 | 8.41 | 6.31 |
| MT | female | English | 17.42 | 14.63 | 11.24 | 8.44 |
| JT | female | Scottish | 7.46 | 6.52 | 4.87 | 3.93 |
| JC | female | Scottish | 7.25 | 4.53 | 5.25 | 2.53 |
| KG | female | Scottish | 11.50 | 10.22 | 8.03 | 6.74 |
| DN | female | Scottish | 8.29 | 5.86 | 6.47 | 4.04 |
| SS | female | Scottish | 16.72 | 11.88 | 11.92 | 7.07 |
| AW | female | Scottish | 15.24 | 12.94 | 10.85 | 8.55 |
| JD | female | Scottish | 11.25 | 9.55 | 6.74 | 5.03 |
| JO | female | Scottish | 12.97 | 11.20 | 9.06 | 7.28 |

**Table A.3.** Span measures for each speaker from the English Speech Database, measured in Semitones

| | | | Span (Hz) | | | Level (Hz) | |
|---|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | meanf0±2sds | 95 - 5%f0 | 90 - 10%f0 | meanf0 | medianf0 |
| AP | male | Scottish | 135.69 | 95.16 | 71.17 | 133.96 | 129.24 |
| HM | male | Scottish | 117.63 | 110.68 | 58.73 | 129.93 | 127.23 |
| JS | male | Scottish | 117.92 | 90.98 | 65.08 | 116.68 | 110.03 |
| AB | male | Scottish | 103.20 | 83.15 | 64.12 | 132.91 | 131.63 |
| GF1 | male | Scottish | 133.34 | 113.44 | 89.97 | 116.08 | 118.03 |
| FH | male | Scottish | 97.61 | 74.26 | 54.15 | 114.24 | 107.75 |
| TM | male | Scottish | 83.59 | 66.69 | 47.11 | 100.21 | 96.82 |
| KW | male | Scottish | 108.12 | 87.00 | 65.42 | 109.73 | 105.95 |
| SM | male | English | 106.36 | 89.26 | 62.35 | 119.73 | 117.00 |
| GF2 | male | English | 83.48 | 64.26 | 34.75 | 113.16 | 110.12 |
| RC | male | English | 99.44 | 71.89 | 54.19 | 112.39 | 106.82 |
| ME | male | English | 111.18 | 90.97 | 57.48 | 117.60 | 113.76 |
| RL | male | English | 97.68 | 73.32 | 52.11 | 113.36 | 109.77 |
| VR | male | English | 109.22 | 86.52 | 67.99 | 100.52 | 94.66 |
| JB | male | English | 154.61 | 132.19 | 106.07 | 131.96 | 129.86 |
| GB | male | English | 117.88 | 100.26 | 74.90 | 124.79 | 121.13 |
| FL | female | English | 145.03 | 137.28 | 95.10 | 175.00 | 179.17 |
| NG | female | English | 148.40 | 142.01 | 64.67 | 202.39 | 200.61 |
| SO | female | English | 200.72 | 159.67 | 94.82 | 208.03 | 201.57 |
| NC | female | English | 210.92 | 205.37 | 118.46 | 188.00 | 184.73 |
| JK | female | English | 167.69 | 154.06 | 84.78 | 203.23 | 204.22 |
| JV | female | English | 157.33 | 143.58 | 118.18 | 180.08 | 185.52 |
| RS | female | English | 156.16 | 125.66 | 84.84 | 167.80 | 165.54 |
| MT | female | English | 195.80 | 151.64 | 113.04 | 173.87 | 162.36 |
| JT | female | Scottish | 148.48 | 134.95 | 62.61 | 218.10 | 219.20 |
| JC | female | Scottish | 158.10 | 144.27 | 109.79 | 195.65 | 201.96 |
| KG | female | Scottish | 165.12 | 122.41 | 90.33 | 194.42 | 189.39 |
| DN | female | Scottish | 127.83 | 127.95 | 54.13 | 192.86 | 196.37 |
| SS | female | Scottish | 214.92 | 205.22 | 114.06 | 199.60 | 197.35 |
| AW | female | Scottish | 210.63 | 156.14 | 113.74 | 213.89 | 206.93 |
| JD | female | Scottish | 164.06 | 158.44 | 77.53 | 183.72 | 182.99 |
| JO | female | Scottish | 168.17 | 126.17 | 97.54 | 182.07 | 178.06 |

**Table A.4.** Long term distributional properties of f0 for each speaker from the English Speech Database, measured in Hz

| | | | Span (ERB) | | | Level (ERB) | |
|---|---|---|---|---|---|---|---|
| Speaker | Sex | Accent | meanf0±2sds | 95 - 5%f0 | 90 - 10%f0 | meanf0 | medianf0 |
| AP | male | Scottish | 4.34 | 3.30 | 2.60 | 4.30 | 4.20 |
| HM | male | Scottish | 3.90 | 3.72 | 2.20 | 4.20 | 4.14 |
| JS | male | Scottish | 3.90 | 3.18 | 2.40 | 3.87 | 3.70 |
| AB | male | Scottish | 3.52 | 2.95 | 2.38 | 4.28 | 4.25 |
| GF1 | male | Scottish | 4.29 | 3.79 | 3.15 | 3.86 | 3.91 |
| FH | male | Scottish | 3.36 | 2.69 | 2.05 | 3.81 | 3.62 |
| TM | male | Scottish | 2.97 | 2.46 | 1.82 | 3.44 | 3.34 |
| KW | male | Scottish | 3.65 | 3.07 | 2.42 | 3.69 | 3.59 |
| SM | male | English | 3.60 | 3.13 | 2.32 | 3.95 | 3.88 |
| GF2 | male | English | 2.96 | 2.38 | 1.38 | 3.78 | 3.70 |
| RC | male | English | 3.41 | 2.62 | 2.06 | 3.76 | 3.61 |
| ME | male | English | 3.73 | 3.18 | 2.16 | 3.90 | 3.80 |
| RL | male | English | 3.37 | 2.66 | 1.99 | 3.79 | 3.69 |
| VR | male | English | 3.68 | 3.05 | 2.50 | 3.44 | 3.28 |
| JB | male | English | 4.79 | 4.26 | 3.59 | 4.25 | 4.20 |
| GB | male | English | 3.90 | 3.44 | 2.71 | 4.08 | 3.99 |
| FL | female | English | 4.57 | 4.38 | 3.29 | 5.23 | 5.32 |
| NG | female | English | 4.64 | 4.50 | 2.39 | 5.80 | 5.76 |
| SO | female | English | 5.76 | 4.90 | 3.29 | 5.91 | 5.78 |
| NC | female | English | 5.96 | 5.85 | 3.92 | 5.51 | 5.44 |
| JK | female | English | 5.08 | 4.77 | 3.00 | 5.81 | 5.83 |
| JV | female | English | 4.85 | 4.53 | 3.91 | 5.34 | 5.46 |
| RS | female | English | 4.82 | 4.10 | 3.00 | 5.08 | 5.03 |
| MT | female | English | 5.66 | 4.72 | 3.78 | 5.21 | 4.96 |
| JT | female | Scottish | 4.65 | 4.33 | 2.33 | 6.10 | 6.12 |
| JC | female | Scottish | 4.85 | 4.55 | 3.69 | 5.66 | 5.79 |
| KG | female | Scottish | 5.02 | 4.02 | 3.16 | 5.64 | 5.53 |
| DN | female | Scottish | 4.15 | 4.16 | 2.05 | 5.61 | 5.68 |
| SS | female | Scottish | 6.04 | 5.85 | 3.80 | 5.74 | 5.70 |
| AW | female | Scottish | 5.95 | 4.82 | 3.80 | 6.02 | 5.88 |
| JD | female | Scottish | 5.00 | 4.87 | 2.79 | 5.42 | 5.40 |
| JO | female | Scottish | 5.09 | 4.11 | 3.36 | 5.38 | 5.30 |

**Table A.5.** Long term distributional properties of f0 for each speaker from the English Speech Database, measured in ERB

| | | | Span (semitones) | | |
|---|---|---|---|---|---|
| Speaker | Sex | Accent | meanf0±2sds | 95 - 5%f0 | 90 - 10%f0 |
| AP | male | Scottish | 19.32 | 12.18 | 9.16 |
| HM | male | Scottish | 16.90 | 16.26 | 7.71 |
| JS | male | Scottish | 19.27 | 12.87 | 9.54 |
| AB | male | Scottish | 14.19 | 10.94 | 8.48 |
| GF1 | male | Scottish | 22.64 | 18.79 | 15.44 |
| FH | male | Scottish | 15.81 | 10.73 | 7.97 |
| TM | male | Scottish | 15.38 | 11.32 | 8.13 |
| KW | male | Scottish | 18.68 | 13.69 | 10.50 |
| SM | male | English | 16.53 | 12.64 | 9.03 |
| GF2 | male | English | 13.40 | 9.78 | 5.19 |
| RC | male | English | 16.45 | 10.47 | 8.15 |
| ME | male | English | 17.78 | 13.76 | 8.40 |
| RL | male | English | 15.96 | 11.06 | 7.87 |
| VR | male | English | 21.08 | 14.37 | 11.43 |
| JB | male | English | 23.24 | 18.65 | 15.11 |
| GB | male | English | 17.76 | 14.05 | 10.47 |
| FL | female | English | 15.27 | 16.75 | 10.34 |
| NG | female | English | 13.31 | 13.65 | 5.38 |
| SO | female | English | 18.22 | 13.20 | 7.73 |
| NC | female | English | 21.96 | 21.49 | 10.83 |
| JK | female | English | 15.19 | 14.92 | 7.22 |
| JV | female | English | 16.22 | 16.89 | 13.48 |
| RS | female | English | 17.45 | 13.90 | 8.72 |
| MT | female | English | 22.07 | 14.39 | 10.99 |
| JT | female | Scottish | 12.28 | 12.07 | 4.87 |
| JC | female | Scottish | 14.84 | 15.72 | 11.21 |
| KG | female | Scottish | 15.70 | 10.86 | 7.92 |
| DN | female | Scottish | 11.93 | 13.62 | 4.80 |
| SS | female | Scottish | 20.84 | 21.23 | 9.90 |
| AW | female | Scottish | 18.61 | 12.37 | 9.09 |
| JD | female | Scottish | 16.63 | 17.47 | 7.15 |
| JO | female | Scottish | 17.30 | 11.54 | 9.11 |

**Table A.6.** Long term distributional properties of f0 for each speaker from the English Speech Database, measured in Semitones

| Speaker | Sex | Accent | skew | kurtosis |
|---|---|---|---|---|
| AP | male | Scottish | 0.680 | 0.749 |
| HM | male | Scottish | 0.662 | 6.089 |
| JS | male | Scottish | 1.731 | 7.229 |
| AB | male | Scottish | 0.255 | 0.624 |
| GF1 | male | Scottish | 0.301 | 1.275 |
| FH | male | Scottish | 1.779 | 11.169 |
| TM | male | Scottish | 1.402 | 5.126 |
| KW | male | Scottish | 1.161 | 0.025 |
| SM | male | English | 0.779 | 1.292 |
| GF2 | male | English | 1.575 | 8.267 |
| RC | male | English | 2.370 | 23.220 |
| ME | male | English | 1.689 | 10.669 |
| RL | male | English | 2.261 | 18.393 |
| VR | male | English | 1.143 | 1.546 |
| JB | male | English | 0.266 | -0.270 |
| GB | male | English | 0.633 | 0.681 |
| FL | female | English | -1.499 | 2.395 |
| NG | female | English | -0.437 | 4.890 |
| SO | female | English | 0.945 | 4.353 |
| NC | female | English | 0.310 | 0.880 |
| JK | female | English | -0.671 | 2.418 |
| JV | female | English | -1.165 | 1.882 |
| RS | female | English | 0.075 | 1.402 |
| MT | female | English | 0.976 | 0.025 |
| JT | female | Scottish | -1.468 | 4.919 |
| JC | female | Scottish | -1.634 | 3.001 |
| KG | female | Scottish | 0.287 | 0.026 |
| DN | female | Scottish | -1.649 | 4.890 |
| SS | female | Scottish | 0.100 | 1.455 |
| AW | female | Scottish | 0.562 | 2.591 |
| JD | female | Scottish | -0.350 | 2.013 |
| JO | female | Scottish | 0.191 | 0.761 |

**Table A.7.** Skew and kurtosis of f0 measurements for each speaker from the English Speech Database

# Appendix B

# Examples of range measurements in Experiment 2

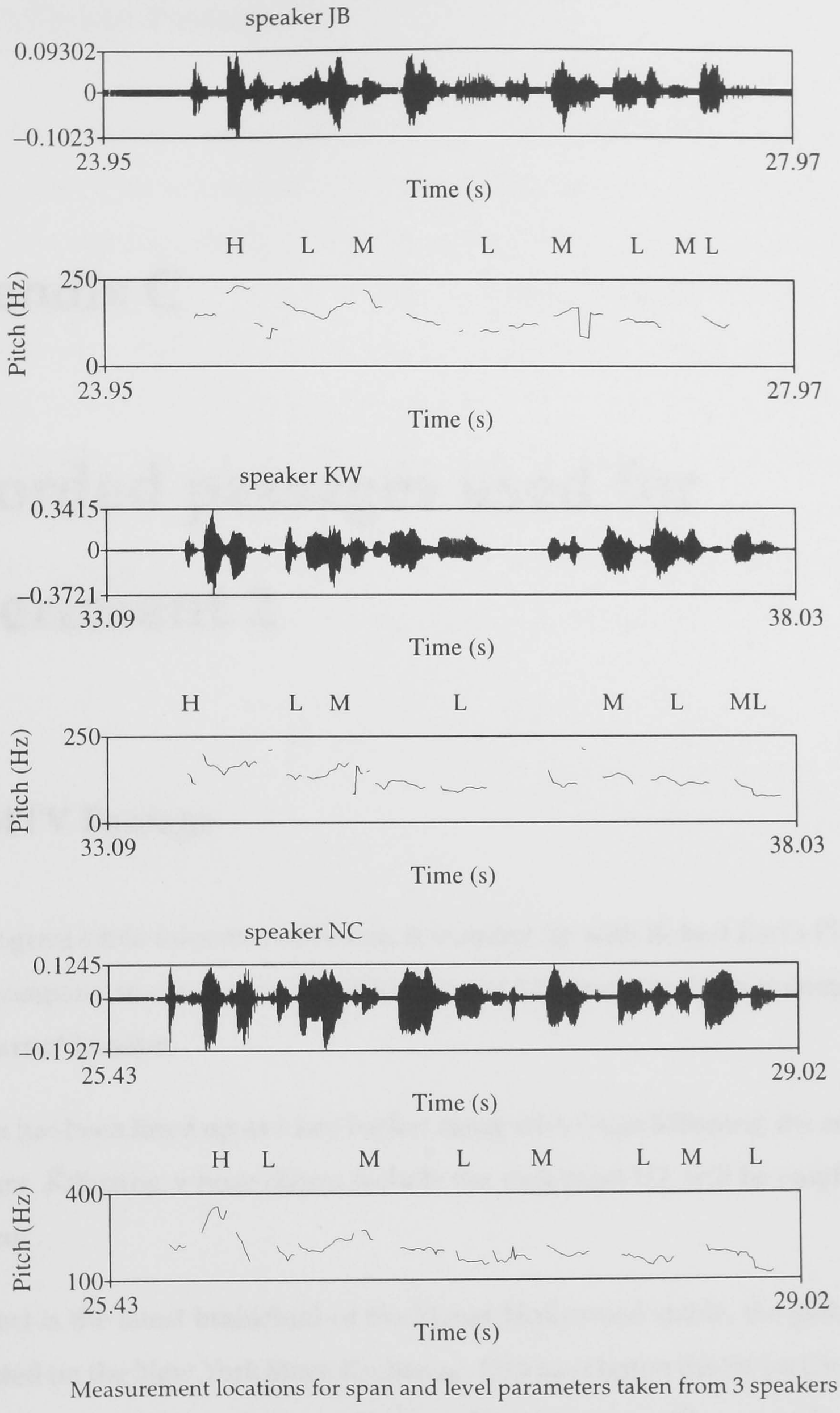
## B.1 Measurements

Figure B.1 shows the waveforms and f0 contours for 3 speakers (JB, KW and NC) saying, *"The project is the latest brainchild of the Planet Hollywood stable"*, with selected measurement points above the f0 contour for each speaker. The sentence-initial high measure is clearly the first peak in the f0 contour. In the examples shown the sentence-initial high is on the f0 peak found on the first accent on the first syllable of *"project"*.

For the M and L measures, the first point to make is that there is an equal amount of measures taken for the three speakers. Apart from clear octave errors in the f0 contour for speaker JB, the f0 contour falls smoothly from the initial H down to the first L. This is not the case for speaker KW. Following the contour from the initial H, there is a fall, a slight rise and then a fall again. The second fall reaches a valley in approximately the same place as the first L measure taken for speaker JB. For the purposes of this experiment, we are interested in consistent measurement points across speakers, so the initial small turning point identifiable for speaker KW was not taken to represent an

L measurement point, nor was the slight peak that followed considered to be the first measurement point for M. Moving on to speaker NC, there is another small turning point identifiable very soon after the sentence-initial high. This is considered to be a segmental perturbation and is not taken as an L measurement point. Clearly such a point is not representative of L as a possible measure of the bottom of the speaker span.

Decisions about further representatives for M and L were made in a similar fashion. Consistent locations of peaks and valleys were sought across speakers, and minor perturbations were ignored for the purposes of experiment 2. The phonological status of the L measures taken (whether they be L*, L- or sag) was not considered. Discussion of this issue of left for section 6.4.




Measurement locations for span and level parameters taken from 3 speakers
(JB, KW and NC). The text for all 3 speakers is
"The project of the latest brainchild of the Planet Hollywood stable"
taken from the MTV passage.

Figure B.1

# Appendix C

# Recorded passages used for Experiment 2

## C.1 MTV Passage

MTV, the giant cable-television network, is teaming up with Robert Earl's Planet Hollywood company, to create a multi-million-pound live-music and entertainment venue in the heart of London.

Madonna has been lined up as a key backer along with Ossie Kilkenny, the accountant to the stars. Kilkenny, whose clients include the rock band U2, will be employed as a consultant.

The project is the latest brainchild of the Planet Hollywood stable, the global leisure group listed on the New York Stock Exchange. Earl has chosen the Swiss Centre in Leicester Square for the huge venue and is thought to have set aside one million pounds for the launch party in December.

Landing MTV as a financial partner is seen within the music industry as a coup.

## C.2 Railways Passage

Everyone knows what is supposed to happen when two Englishmen who have never met before come face to face in a railway compartment - they start talking about the weather. In some cases this may simply be because they happen to find the subject interesting. Most people, though, are not particularly interested in analyses of climatic conditions, so there must be other reasons for conversations of this kind. One explanation is that it can often be quite embarrassing to be alone in the company of someone you are not acquainted with and not speak to them. If no conversation takes place the atmosphere can become rather strained. However, by talking to the other person about about some neutral topic like the weather it is possible to strike up a relationship with him without actually having to say very much.

# Appendix D

# Cutoff levels used for low pass filtering

| Speaker | Sex | Accent | Highest f0 | Cutoff level |
|---|---|---|---|---|
| AP | male | Scottish | 205 | 258.30 |
| HM | male | Scottish | 215 | 270.90 |
| JS | male | Scottish | 235 | 296.10 |
| AB | male | Scottish | 200 | 252.00 |
| GF1 | male | Scottish | 220 | 277.20 |
| FH | male | Scottish | 195 | 245.70 |
| TM | male | Scottish | 180 | 226.80 |
| KW | male | Scottish | 190 | 239.40 |
| SM | male | English | 195 | 245.70 |
| GF2 | male | English | 138 | 173.88 |
| RC | male | English | 220 | 277.20 |
| ME | male | English | 210 | 264.60 |
| RL | male | English | 185 | 233.10 |
| VR | male | English | 205 | 258.30 |
| JB | male | English | 250 | 315.00 |
| GB | male | English | 230 | 289.80 |
| FL | female | English | 240 | 302.50 |
| NG | female | English | 340 | 428.40 |
| SO | female | English | 440 | 500.00 |
| NC | female | English | 350 | 441.00 |
| JK | female | English | 330 | 415.80 |
| JV | female | English | 350 | 441.00 |
| RS | female | English | 310 | 390.60 |
| MT | female | English | 330 | 415.80 |
| JT | female | Scottish | 315 | 396.90 |
| JC | female | Scottish | 280 | 352.80 |
| KG | female | Scottish | 320 | 403.20 |
| DN | female | Scottish | 275 | 346.50 |
| SS | female | Scottish | 355 | 447.30 |
| AW | female | Scottish | 380 | 478.80 |
| JD | female | Scottish | 300 | 378.00 |
| JV | female | Scottish | 260 | 327.60 |

**Table D.1.** Cutoff levels used for low pass filtering used in Experiment 2

# Appendix E

# Results for Experiment 2: Modes

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 5 | 4 | 4 | 6 | 3 | 4 | 5 | 5 |
| tense | 2 | 3 | 2 | 1 | 6 | 1 | 2 | 4 |
| harsh | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 |
| expressive | 3 | 2 | 3 | 3 | 4 | 3 | 5 | 3 |
| deep | 4 | 3 | 5 | 3 | 3 | 3 | 6 | 5 |
| weak | 2 | 5 | 3 | 2 | 2 | 1 | 2 | 1 |
| irritated | 3 | 2 | 1 | 2 | 2 | 2 | 1 | 2 |
| happy | 1 | 2 | 5 | 2 | 2 | 3 | 2 | 2 |
| afraid | 2 | 1 | 3 | 4 | 1 | 1 | 2 | 1 |
| relaxed | 2 | 2 | 2 | 1 | 1 | 6 | 5 | 3 |
| emphatic | 3 | 2 | 5 | 3 | 6 | 2 | 5 | 3 |
| bored | 5 | 3 | 2 | 1 | 2 | 3 | 2 | 3 |
| | English Men | | | | | | | |
| | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 6 | 5 | 3 | 7 | 3 | 4 | 6 | 5 |
| tense | 1 | 5 | 5 | 1 | 5 | 1 | 2 | 2 |
| harsh | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| expressive | 5 | 2 | 3 | 5 | 5 | 5 | 6 | 3 |
| deep | 4 | 5 | 3 | 5 | 3 | 5 | 3 | 4 |
| weak | 2 | 3 | 5 | 1 | 6 | 1 | 2 | 2 |
| irritated | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| happy | 5 | 2 | 2 | 3 | 3 | 4 | 2 | 2 |
| afraid | 1 | 1 | 5 | 1 | 3 | 1 | 1 | 1 |
| relaxed | 5 | 3 | 2 | 5 | 2 | 5 | 5 | 5 |
| emphatic | 5 | 3 | 3 | 2 | 4 | 2 | 5 | 5 |
| bored | 3 | 3 | 3 | 2 | 2 | 1 | 2 | 1 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 3 | 4 | 3 | 5 | 5 | 3 | 5 | 7 |
| tense | 5 | 2 | 4 | 1 | 5 | 4 | 5 | 3 |
| harsh | 4 | 1 | 5 | 3 | 2 | 2 | 2 | 4 |
| expressive | 1 | 3 | 4 | 5 | 2 | 2 | 4 | 3 |
| deep | 3 | 1 | 3 | 2 | 2 | 2 | 3 | 3 |
| weak | 3 | 3 | 2 | 1 | 3 | 4 | 2 | 2 |
| irritated | 4 | 1 | 1 | 1 | 4 | 5 | 3 | 2 |
| happy | 1 | 5 | 4 | 5 | 1 | 1 | 2 | 3 |
| afraid | 3 | 1 | 2 | 1 | 1 | 1 | 3 | 1 |
| relaxed | 1 | 5 | 2 | 5 | 3 | 2 | 2 | 2 |
| emphatic | 1 | 3 | 5 | 4 | 4 | 1 | 3 | 6 |
| bored | 6 | 2 | 2 | 2 | 3 | 2 | 2 | 1 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 4 | 2 | 4 | 3 | 5 | 7 | 5 | 5 |
| tense | 2 | 2 | 4 | 5 | 2 | 2 | 5 | 3 |
| harsh | 1 | 1 | 2 | 2 | 1 | 4 | 2 | 1 |
| expressive | 4 | 2 | 5 | 3 | 6 | 4 | 4 | 2 |
| deep | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 3 |
| weak | 3 | 2 | 2 | 5 | 2 | 1 | 3 | 1 |
| irritated | 2 | 4 | 1 | 2 | 1 | 1 | 3 | 1 |
| happy | 5 | 2 | 5 | 2 | 5 | 3 | 2 | 2 |
| afraid | 5 | 1 | 3 | 2 | 2 | 1 | 5 | 2 |
| relaxed | 2 | 1 | 2 | 1 | 4 | 3 | 2 | 2 |
| emphatic | 4 | 2 | 5 | 2 | 5 | 5 | 3 | 4 |
| bored | 2 | 7 | 3 | 3 | 3 | 3 | 4 | 2 |

**Table E.1.** Mode results of normal speech for all speakers reading the MTV passage

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 5 | 3 | 5 | 3 | 2 | 5 | 5 | 5 |
| tense | 2 | 5 | 3 | 5 | 2 | 2 | 2 | 3 |
| harsh | 1 | 1 | 5 | 2 | 2 | 1 | 4 | 2 |
| expressive | 3 | 3 | 5 | 4 | 3 | 4 | 2 | 4 |
| deep | 4 | 5 | 5 | 6 | 5 | 4 | 6 | 7 |
| weak | 1 | 2 | 1 | 3 | 3 | 2 | 1 | 1 |
| irritated | 2 | 1 | 2 | 3 | 3 | 1 | 3 | 2 |
| happy | 1 | 2 | 5 | 2 | 1 | 2 | 1 | 2 |
| afraid | 2 | 5 | 1 | 1 | 2 | 2 | 1 | 2 |
| relaxed | 5 | 3 | 6 | 2 | 3 | 5 | 2 | 4 |
| emphatic | 3 | 3 | 5 | 4 | 3 | 4 | 3 | 5 |
| bored | 2 | 6 | 1 | 3 | 1 | 4 | 4 | 2 |
| | English Men | | | | | | | |
| | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 3 | 2 | 4 | 5 | 6 | 5 | 5 | 5 |
| tense | 2 | 2 | 2 | 2 | 4 | 3 | 5 | 2 |
| harsh | 3 | 1 | 2 | 2 | 2 | 1 | 3 | 2 |
| expressive | 5 | 2 | 2 | 5 | 2 | 3 | 5 | 5 |
| deep | 5 | 5 | 4 | 6 | 4 | 2 | 2 | 5 |
| weak | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 3 |
| irritated | 2 | 5 | 3 | 3 | 3 | 1 | 3 | 2 |
| happy | 3 | 1 | 2 | 2 | 4 | 3 | 3 | 3 |
| afraid | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 |
| relaxed | 5 | 5 | 5 | 4 | 4 | 5 | 3 | 6 |
| emphatic | 5 | 3 | 2 | 5 | 2 | 4 | 3 | 6 |
| bored | 1 | 5 | 3 | 3 | 5 | 3 | 3 | 3 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 2 | 3 | 5 | 5 | 2 | 2 | 3 | 5 |
| tense | 3 | 5 | 5 | 2 | 5 | 5 | 3 | 2 |
| harsh | 2 | 1 | 5 | 2 | 1 | 2 | 1 | 2 |
| expressive | 2 | 3 | 4 | 5 | 3 | 3 | 4 | 5 |
| deep | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
| weak | 4 | 5 | 3 | 2 | 3 | 5 | 4 | 2 |
| irritated | 4 | 3 | 3 | 2 | 3 | 2 | 6 | 2 |
| happy | 1 | 1 | 2 | 4 | 2 | 1 | 2 | 5 |
| afraid | 2 | 4 | 3 | 2 | 3 | 4 | 1 | 1 |
| relaxed | 3 | 2 | 2 | 5 | 2 | 3 | 3 | 5 |
| emphatic | 2 | 2 | 4 | 6 | 3 | 4 | 2 | 5 |
| bored | 3 | 3 | 2 | 2 | 2 | 5 | 5 | 2 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 3 | 2 | 5 | 5 | 5 | 5 | 5 | 5 |
| tense | 5 | 5 | 2 | 3 | 2 | 3 | 5 | 2 |
| harsh | 2 | 1 | 1 | 3 | 1 | 2 | 5 | 3 |
| expressive | 2 | 2 | 4 | 2 | 5 | 4 | 2 | 5 |
| deep | 1 | 1 | 2 | 2 | 1 | 3 | 4 | 3 |
| weak | 6 | 1 | 1 | 2 | 2 | 2 | 3 | 2 |
| irritated | 2 | 2 | 3 | 2 | 1 | 3 | 4 | 2 |
| happy | 2 | 1 | 3 | 2 | 3 | 4 | 2 | 2 |
| afraid | 3 | 5 | 1 | 3 | 1 | 2 | 1 | 2 |
| relaxed | 2 | 1 | 3 | 2 | 5 | 3 | 1 | 5 |
| emphatic | 4 | 2 | 2 | 3 | 5 | 4 | 2 | 4 |
| bored | 1 | 3 | 5 | 3 | 2 | 3 | 3 | 5 |

**Table E.2.** Mode results of low pass filtered speech for all speakers reading the MTV passage

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 5 | 2 | 4 | 3 | 5 | 5 | 6 | 5 |
| tense | 1 | 4 | 1 | 5 | 2 | 1 | 1 | 3 |
| harsh | 2 | 1 | 1 | 3 | 2 | 1 | 1 | 2 |
| expressive | 5 | 1 | 3 | 4 | 6 | 5 | 6 | 3 |
| deep | 4 | 3 | 5 | 3 | 3 | 4 | 4 | 3 |
| weak | 2 | 3 | 1 | 2 | 2 | 1 | 1 | 1 |
| irritated | 2 | 2 | 1 | 3 | 1 | 1 | 3 | 2 |
| happy | 5 | 1 | 5 | 3 | 3 | 3 | 2 | 2 |
| afraid | 2 | 3 | 1 | 3 | 1 | 1 | 1 | 1 |
| relaxed | 5 | 1 | 5 | 2 | 7 | 5 | 4 | 3 |
| emphatic | 5 | 1 | 5 | 4 | 6 | 5 | 5 | 3 |
| bored | 3 | 3 | 1 | 2 | 1 | 2 | 2 | 6 |
| | English Men | | | | | | | |
| | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 5 | 3 | 2 | 6 | 5 | 6 | 7 | 5 |
| tense | 3 | 4 | 3 | 2 | 3 | 1 | 1 | 1 |
| harsh | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| expressive | 5 | 1 | 5 | 5 | 6 | 6 | 6 | 5 |
| deep | 2 | 5 | 2 | 4 | 3 | 5 | 5 | 3 |
| weak | 2 | 4 | 2 | 1 | 3 | 1 | 1 | 1 |
| irritated | 1 | 6 | 1 | 1 | 2 | 1 | 1 | 1 |
| happy | 5 | 1 | 2 | 2 | 5 | 5 | 5 | 3 |
| afraid | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 |
| relaxed | 3 | 1 | 1 | 5 | 2 | 6 | 5 | 6 |
| emphatic | 3 | 1 | 3 | 4 | 6 | 5 | 7 | 5 |
| bored | 2 | 7 | 3 | 2 | 1 | 1 | 1 | 1 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 1 | 5 | 7 | 6 | 6 | 2 | 4 | 7 |
| tense | 7 | 3 | 1 | 1 | 3 | 5 | 5 | 1 |
| harsh | 2 | 2 | 2 | 1 | 2 | 3 | 1 | 1 |
| expressive | 2 | 5 | 6 | 6 | 4 | 2 | 4 | 6 |
| deep | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 |
| weak | 3 | 2 | 1 | 1 | 2 | 3 | 2 | 1 |
| irritated | 3 | 1 | 1 | 1 | 2 | 3 | 1 | 4 |
| happy | 2 | 3 | 7 | 4 | 2 | 1 | 2 | 4 |
| afraid | 5 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| relaxed | 1 | 3 | 6 | 4 | 2 | 2 | 3 | 6 |
| emphatic | 2 | 2 | 5 | 5 | 6 | 2 | 5 | 5 |
| bored | 3 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 2 | 2 | 3 | 5 | 5 | 5 | 3 | 5 |
| tense | 6 | 5 | 4 | 4 | 2 | 2 | 5 | 3 |
| harsh | 3 | 2 | 2 | 1 | 3 | 2 | 2 | 1 |
| expressive | 3 | 2 | 4 | 2 | 5 | 4 | 4 | 5 |
| deep | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 |
| weak | 2 | 3 | 3 | 3 | 1 | 3 | 5 | 1 |
| irritated | 2 | 3 | 1 | 2 | 1 | 3 | 2 | 1 |
| happy | 3 | 1 | 4 | 2 | 5 | 2 | 1 | 5 |
| afraid | 2 | 5 | 4 | 2 | 1 | 2 | 6 | 1 |
| relaxed | 2 | 1 | 2 | 2 | 5 | 5 | 2 | 6 |
| emphatic | 2 | 1 | 5 | 2 | 5 | 4 | 2 | 5 |
| bored | 2 | 7 | 2 | 5 | 1 | 2 | 3 | 1 |

**Table E.3.** Mode results of normal speech for all speakers reading the Railways passage

| feature | Scottish Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AP | HM | JS | AB | GF1 | FH | TM | KW |
| confident | 4 | 3 | 5 | 5 | 2 | 5 | 5 | 5 |
| tense | 1 | 4 | 2 | 2 | 3 | 2 | 3 | 2 |
| harsh | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| expressive | 6 | 3 | 5 | 4 | 2 | 5 | 4 | 6 |
| deep | 5 | 6 | 3 | 4 | 6 | 6 | 7 | 7 |
| weak | 1 | 2 | 1 | 2 | 5 | 1 | 1 | 2 |
| irritated | 1 | 3 | 1 | 1 | 2 | 2 | 3 | 1 |
| happy | 6 | 3 | 4 | 2 | 1 | 5 | 2 | 5 |
| afraid | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 |
| relaxed | 3 | 3 | 6 | 5 | 3 | 6 | 2 | 6 |
| emphatic | 4 | 4 | 6 | 4 | 2 | 5 | 4 | 6 |
| bored | 2 | 3 | 2 | 5 | 4 | 2 | 3 | 2 |
| | English Men | | | | | | | |
| | SM | GF2 | RC | ME | RL | VR | JB | GB |
| confident | 5 | 3 | 6 | 5 | 5 | 5 | 5 | 4 |
| tense | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| harsh | 2 | 2 | 2 | 1 | 3 | 2 | 1 | 1 |
| expressive | 5 | 2 | 4 | 5 | 4 | 5 | 5 | 4 |
| deep | 5 | 7 | 5 | 5 | 6 | 6 | 3 | 3 |
| weak | 1 | 4 | 2 | 2 | 2 | 2 | 1 | 2 |
| irritated | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 |
| happy | 2 | 2 | 4 | 3 | 4 | 5 | 3 | 3 |
| afraid | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 1 |
| relaxed | 3 | 5 | 6 | 5 | 4 | 5 | 4 | 5 |
| emphatic | 4 | 1 | 6 | 5 | 4 | 6 | 5 | 5 |
| bored | 2 | 6 | 1 | 3 | 1 | 1 | 3 | 3 |
| | English Women | | | | | | | |
| | FL | NG | SO | NC | JK | JV | RS | MT |
| confident | 4 | 3 | 5 | 6 | 5 | 2 | 5 | 6 |
| tense | 5 | 3 | 2 | 2 | 3 | 5 | 2 | 2 |
| harsh | 1 | 2 | 2 | 1 | 5 | 4 | 2 | 2 |
| expressive | 4 | 3 | 4 | 6 | 3 | 3 | 4 | 6 |
| deep | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 3 |
| weak | 5 | 5 | 2 | 1 | 2 | 5 | 1 | 2 |
| irritated | 3 | 2 | 1 | 1 | 4 | 3 | 1 | 1 |
| happy | 2 | 2 | 4 | 5 | 2 | 2 | 1 | 3 |
| afraid | 5 | 2 | 1 | 1 | 1 | 3 | 1 | 2 |
| relaxed | 2 | 2 | 5 | 5 | 2 | 2 | 5 | 5 |
| emphatic | 2 | 2 | 5 | 5 | 5 | 3 | 3 | 6 |
| bored | 6 | 3 | 1 | 1 | 1 | 3 | 1 | 1 |
| | Scottish Women | | | | | | | |
| | JT | JC | KG | DN | SS | AW | JD | JO |
| confident | 2 | 4 | 4 | 5 | 5 | 5 | 5 | 5 |
| tense | 5 | 5 | 3 | 2 | 1 | 2 | 2 | 2 |
| harsh | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 1 |
| expressive | 3 | 2 | 5 | 4 | 6 | 7 | 5 | 6 |
| deep | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 1 |
| weak | 1 | 6 | 3 | 2 | 1 | 1 | 2 | 2 |
| irritated | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 |
| happy | 2 | 1 | 1 | 2 | 5 | 6 | 1 | 4 |
| afraid | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 2 |
| relaxed | 3 | 2 | 4 | 2 | 4 | 4 | 3 | 5 |
| emphatic | 4 | 4 | 3 | 3 | 6 | 6 | 5 | 5 |
| bored | 4 | 3 | 2 | 5 | 1 | 1 | 1 | 4 |

**Table E.4.** Mode results of low pass filtered speech for all speakers reading the Railways passage

# Appendix F

# Results of correlation analyses for Experiment 2

| | Normal Speech | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level | | | | | | Span | | | | | |
| Feature | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | -0.187 | | -0.279 | | **0.464** | √ | **0.403** | | **0.439** | | **0.456** | |
| tense | 0.169 | | 0.274 | | -0.158 | | -0.121 | | -0.115 | | -0.141 | |
| harsh | 0.184 | | 0.199 | | **0.307** | | **0.358** | √ | **0.316** | | **0.322** | |
| expressive | -0.212 | | -0.229 | | **0.355** | √ | 0.225 | | **0.339** | | **0.313** | |
| deep | **−0.834** | √ | **−0.802** | | -0.228 | | −0.331 | √ | -0.277 | | -0.310 | |
| weak | -0.241 | | 0.275 | | **−0.470** | √ | **−0.443** | | **−0.359** | | **−0.381** | |
| irritated | -0.294 | | 0.246 | | **−0.400** | √ | -0.194 | | **−0.350** | | **−0.314** | |
| happy | -0.029 | | 0.077 | | **0.314** | | 0.158 | | **0.363** | √ | **0.328** | |
| afraid | 0.074 | | 0.230 | | 0.022 | | -0.119 | | 0.088 | | -0.010 | |
| relaxed | **−0.309** | | **−0.427** | √ | 0.143 | | 0.083 | | 0.135 | | 0.146 | |
| emphatic | -0.045 | | -0.114 | | **0.558** | | **0.438** | | **0.580** | √ | **0.549** | |
| bored | 0.255 | | **0.305** | √ | -0.276 | | -0.202 | | -0.204 | | -0.212 | |
| | Filtered Speech | | | | | | | | | | | |
| | Level | | | | | | Span | | | | | |
| | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | **0.328** | √ | 0.287 | | **0.486** | √ | **0.343** | | **0.393** | | **0.332** | |
| tense | **0.487** | √ | **0.408** | | -0.113 | | 0.055 | | -0.106 | | -0.028 | |
| harsh | -0.125 | | -0.117 | | 0.076 | | 0.016 | | 0.120 | | 0.066 | |
| expressive | -0.138 | | **−0.297** | √ | **0.647** | √ | **0.582** | | **0.605** | | **0.635** | |
| deep | **−0.772** | √ | **−0.681** | | **−0.311** | | **−0.419** | √ | **−0.338** | | **−0.376** | |
| weak | **0.426** | √ | **0.386** | | -0.007 | | 0.029 | | 0.113 | | 0.133 | |
| irritated | 0.076 | | 0.211 | | -0.109 | | -0.185 | | 0.031 | | -0.15 | |
| happy | -0.168 | | -0.235 | | **0.614** | √ | **0.520** | | **0.514** | | **0.497** | |
| afraid | **0.542** | √ | **0.436** | | -0.150 | | 0.019 | | -0.124 | | -0.072 | |
| relaxed | **−0.553** | | **−0.589** | √ | 0.178 | | 0.057 | | 0.117 | | 0.071 | |
| emphatic | -0.209 | | **−0.360** | √ | **0.384** | | **0.389** | √ | 0.285 | | **0.319** | |
| bored | -0.076 | | 0.049 | | -0.288 | | -0.284 | | -0.322 | | **−0.337** | √ |

**Table F.1.** Results of correlation analyses for 2 linguistic measures of level and 4 linguistic measures of span (measured in Hz) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level | | | | Span | | | | | | | |
| Feature | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | -0.197 | | **−0.317** | √ | **0.485** | √ | **0.439** | | **0.432** | | **0.454** | |
| tense | **0.441** | | **0.533** | √ | **−0.425** | √ | **−0.311** | | **−0.380** | | **−0.356** | |
| harsh | **0.401** | √ | **0.356** | | -0.025 | | 0.065 | | 0.082 | | 0.131 | |
| expressive | **−0.324** | | **0.360** | √ | **0.363** | | 0.216 | | **0.403** | √ | **0.359** | |
| deep | **−0.718** | | **−0.745** | √ | -0.217 | | **−0.311** | | **−0.328** | | **−0.331** | √ |
| weak | **0.373** | | **0.518** | √ | **−0.436** | √ | **−0.349** | | **−0.350** | | **−0.358** | |
| irritated | 0.189 | | 0.239 | | **−0.352** | | -0.183 | | **−0.397** | √ | **−0.383** | |
| happy | -0.128 | | -0.214 | | **0.493** | √ | **0.364** | | **0.476** | | **0.450** | |
| afraid | **0.409** | | **0.536** | √ | -0.273 | | -0.210 | | -0.188 | | -0.193 | |
| relaxed | -0.253 | | **−0.340** | √ | **0.585** | √ | **0.492** | | **0.541** | | **0.539** | |
| emphatic | -0.289 | | **−0.384** | √ | **0.554** | √ | **0.444** | | **0.477** | | **0.452** | |
| bored | 0.191 | | 0.264 | | **−0.492** | √ | **−0.346** | | **−0.479** | | **−0.463** | |
| | Filtered Speech | | | | | | | | | | | |
| | Level | | | | Span | | | | | | | |
| | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | **−0.298** | | **−0.301** | √ | **0.421** | √ | **0.316** | | **0.381** | | **0.355** | |
| tense | **0.315** | | **0.369** | √ | **−0.522** | √ | **−0.409** | | **−0.432** | | **−0.412** | |
| harsh | -0.009 | | 0.022 | | **−0.385** | √ | -0.276 | | **−0.367** | | **0.366** | |
| expressive | -0.222 | | -0.235 | | **0.635** | √ | **0.545** | | **0.555** | | **0.516** | |
| deep | **−0.823** | √ | **−0.737** | | **−0.462** | | **−0.535** | √ | **−0.501** | | **−0.526** | |
| weak | 0.202 | | **0.313** | √ | **−0.422** | √ | **−0.300** | | **−0.382** | | **−0.317** | |
| irritated | 0.065 | | 0.015 | | **−0.528** | √ | **−0.377** | | **−0.512** | | **−0.442** | |
| happy | **−0.308** | | **−0.364** | √ | **0.366** | √ | **0.299** | | 0.272 | | 0.242 | |
| afraid | 0.249 | | 0.253 | | **−0.402** | √ | -0.277 | | **−0.369** | | **−0.304** | |
| relaxed | **−0.547** | √ | **−0.486** | | **0.341** | √ | 0.176 | | 0.270 | | 0.215 | |
| emphatic | -0.270 | | **−0.347** | √ | **0.545** | √ | **0.492** | | **0.462** | | **0.461** | |
| bored | 0.048 | | 0.050 | | **−0.497** | | **−0.441** | | **−0.545** | √ | **−0.509** | |

**Table F.2.** Results of correlation analyses for 2 linguistic measures of level and 4 linguistic measures of span (measured in Hz) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level | | | | | Span | | | | | | |
| Feature | L | | F | | | M-L | | M-F | | H-L | | H-F |
| confident | -0.187 | | -0.279 | | | **0.512** | | **0.528** | √ | **0.464** | | **0.507** |
| tense | 0.169 | | 0.274 | | | -0.272 | | -0.257 | | -0.203 | | -0.227 |
| harsh | 0.184 | | 0.199 | | | 0.269 | | 0.270 | | **0.305** | √ | 0.285 |
| expressive | -0.212 | | -0.299 | | | **0.411** | √ | **0.350** | | **0.396** | | **0.402** |
| deep | **−0.834** | √ | **0.802** | | | 0.024 | | 0.001 | | -0.041 | | -0.027 |
| weak | 0.241 | | 0.275 | | | **−0.583** | √ | **−0.580** | | **−0.485** | | **−0.523** |
| irritated | 0.294 | | 0.246 | | | **−0.516** | √ | **−0.353** | | **−0.487** | | **−0.430** |
| happy | 0.029 | | 0.077 | | | 0.292 | | 0.154 | | **0.343** | √ | 0.278 |
| afraid | 0.074 | | -0.230 | | | -0.063 | | -0.235 | | 0.003 | | -0.122 |
| relaxed | **−0.309** | | **−0.427** | √ | | 0.245 | | 0.278 | | 0.169 | | 0.287 |
| emphatic | 0.045 | | -0.114 | | | **0.564** | | **0.496** | | **0.603** | √ | **0.585** |
| bored | 0.255 | | **0.305** | √ | | **−0.346** | √ | -0.273 | | -0.258 | | -0.288 |
| | Filtered Speech | | | | | | | | | | | |
| | Level | | | | | Span | | | | | | |
| | L | | F | | | M-L | | M-F | | H-L | | H-F |
| confident | **−0.328** | √ | -0.287 | | | **0.598** | √ | **0.517** | | **0.529** | | **0.525** |
| tense | **0.487** | √ | **0.408** | | | -0.264 | | -0.166 | | -0.295 | | -0.223 |
| harsh | -0.125 | | -0.117 | | | 0.103 | | 0.077 | | 0.163 | | 0.100 |
| expressive | -0.138 | | **−0.297** | √ | | **0.734** | √ | **0.713** | | **0.718** | | **0.731** |
| deep | **−0.772** | √ | **−0.681** | | | -0.085 | | -0.078 | | -0.146 | | -0.189 |
| weak | **0.426** | √ | **0.386** | | | -0.185 | | -0.190 | | -0.098 | | -0.098 |
| irritated | 0.070 | | 0.211 | | | -0.188 | | -0.282 | | -0.086 | | -0.161 |
| happy | -0.168 | | -0.235 | | | **0.692** | √ | **0.620** | | **0.604** | | **0.625** |
| afraid | **0.542** | √ | **0.436** | | | -0.282 | | -0.179 | | -0.259 | | -0.221 |
| relaxed | **−0.553** | | **−0.589** | √ | | **0.392** | √ | **0.318** | | **0.341** | | **0.337** |
| emphatic | -0.209 | | **−0.360** | √ | | **0.480** | | **0.548** | √ | **0.410** | | **0.469** |
| bored | -0.076 | | 0.049 | | | -0.235 | | **−0.301** | | **−0.310** | | **−0.341** √ |

**Table F.3.** Results of correlation analyses for 2 linguistic measures of level and 4 linguistic measures of span (measured in ERB) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| Feature | Normal Speech | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level | | | | Span | | | | | | | |
| | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | -0.197 | | **−0.317** | √ | **0.532** | | **0.594** | | **0.514** | | **0.627** | √ |
| tense | **0.441** | | **0.533** | √ | **−0.597** | | **−0.584** | | **−0.576** | | **−0.622** | √ |
| harsh | **0.401** | √ | **0.356** | | -0.135 | | -0.074 | | -0.041 | | -0.051 | |
| expressive | **−0.324** | | **−0.360** | √ | **0.473** | | **0.401** | | **0.510** | | **0.536** | √ |
| deep | **−0.718** | | **−0.745** | √ | -0.012 | | 0.015 | | -0.131 | | -0.079 | |
| weak | **0.373** | | **0.518** | √ | **−0.568** | | **−0.556** | | **−0.484** | | **−0.569** | √ |
| irritated | 0.189 | | 0.239 | | **−0.377** | | -0.290 | | **−0.446** | √ | **−0.429** | |
| happy | -0.128 | | -0.214 | | **0.548** | | **0.455** | | **0.568** | | **0.572** | √ |
| afraid | **0.409** | | **0.536** | √ | **−0.368** | | **−0.418** | √ | **−0.297** | | **−0.410** | |
| relaxed | -0.253 | | **−0.340** | √ | **0.670** | | **0.658** | | **0.679** | | **0.698** | √ |
| emphatic | -0.289 | | **−0.384** | √ | **0.638** | √ | **0.598** | | **0.578** | | **0.635** | |
| bored | 0.191 | | 0.264 | | **−0.583** | | **−0.478** | | **−0.584** | √ | **−0.573** | |
| | Filtered Speech | | | | | | | | | | | |
| | Level | | | | Span | | | | | | | |
| | L | | F | | M-L | | M-F | | H-L | | H-F | |
| confident | **−0.298** | | **−0.301** | √ | **0.499** | | **0.463** | | **0.467** | | **0.508** | √ |
| tense | **0.315** | | **0.369** | √ | **−0.612** | √ | **−0.597** | | **−0.562** | | **−0.609** | |
| harsh | -0.009 | | 0.022 | | **−0.399** | √ | **−0.321** | | **−0.383** | | **−0.325** | |
| expressive | -0.222 | | -0.235 | | **0.741** | √ | **0.688** | | **0.681** | | **0.625** | |
| deep | **−0.823** | √ | **−0.737** | | -0.202 | | -0.169 | | -0.276 | | -0.262 | |
| weak | 0.202 | | **0.313** | √ | **−0.462** | √ | **−0.429** | | **−0.407** | | **−0.424** | |
| irritated | 0.065 | | 0.015 | | **−0.512** | | **−0.395** | | **−0.526** | √ | **−0.456** | |
| happy | **−0.308** | | **−0.364** | √ | **0.528** | √ | **0.514** | | **0.428** | | **0.479** | |
| afraid | 0.249 | | 0.253 | | **−0.407** | | **−0.385** | | **−0.354** | | **−0.420** | √ |
| relaxed | **−0.547** | √ | **−0.486** | | **0.511** | √ | **0.382** | | **0.430** | | **0.405** | |
| emphatic | 0.270 | | **−0.347** | √ | **0.664** | | **0.671** | √ | **0.580** | | **0.657** | |
| bored | 0.048 | | 0.050 | | **−0.500** | | **−0.483** | | **−0.549** | | **−0.595** | √ |

**Table F.4.** Results of correlation analyses for 2 linguistic measures of level and 4 linguistic measures of span (measured in ERB) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech Span | | | | Filtered Speech Span | | | |
|---|---|---|---|---|---|---|---|---|
| Feature | M-L | M-F | H-L | H-F | M-L | M-F | H-L | H-F |
| confident | **0.529** | **0.536** √ | **0.441** | **0.491** | **0.609** √ | **0.506** | **0.594** | **0.529** |
| tense | **−0.410** | **−0.408** | **−0.408** | **−0.426** √ | **−0.359** | -0.258 | **−0.424** √ | **−0.325** |
| harsh | 0.114 | 0.099 | 0.222 | 0.184 | 0.118 | 0.069 | 0.203 | 0.124 |
| expressive | **0.413** | **0.364** | **0.463** √ | **0.434** | **0.687** | **0.698** | **0.699** | **0.738** √ |
| deep | **0.333** √ | 0.295 | 0.254 | **0.306** | 0.196 | 0.171 | 0.077 | 0.102 |
| weak | **−0.632** √ | **−0.629** | **−0.578** | **−0.600** | **−0.309** √ | -0.292 | 0.216 | -0.208 |
| irritated | **−0.546** | **−0.383** | **−0.556** √ | **−0.452** | -0.258 | **−0.351** | -0.228 | **−0.355** √ |
| happy | 0.237 | 0.113 | **0.410** √ | 0.274 | **0.640** √ | **0.558** | **0.632** | **0.607** |
| afraid | -0.160 | **−0.322** √ | -0.137 | **−0.309** | **−0.389** √ | -0.272 | **−0.344** | -0.267 |
| relaxed | **0.341** | **0.396** √ | **0.320** | **0.421** | **0.545** | **0.456** | **0.554** √ | **0.532** |
| emphatic | **0.465** | **0.437** | **0.528** √ | **0.483** | **0.543** | **0.609** | **0.585** | **0.663** √ |
| bored | **−0.453** √ | **−0.387** | **−0.395** | **−0.423** | -0.249 | **−0.341** | **−0.403** | **−0.440** √ |

**Table F.5.** Results of correlation analyses for 4 linguistic measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech Span | | | | Filtered Speech Span | | | |
|---|---|---|---|---|---|---|---|---|
| Feature | M-L | M-F | H-L | H-F | M-L | M-F | H-L | H-F |
| confident | **0.556** | **0.644** | **0.604** | **0.675** √ | **0.491** | **0.473** | **0.542** √ | **0.516** |
| tense | **−0.714** | **−0.721** | **−0.790** √ | **−0.786** | **−0.663** √ | **−0.605** | **−0.636** | **−0.627** |
| harsh | -0.242 | -0.152 | -0.151 | -0.101 | **−0.365** √ | -0.281 | **−0.360** | **−0.305** |
| expressive | **0.520** | **0.509** | **0.643** √ | **0.618** | **0.714** √ | **0.641** | **0.694** | **0.652** |
| deep | **0.308** √ | **0.303** | 0.141 | 0.224 | 0.140 | 0.136 | 0.060 | 0.143 |
| weak | **−0.663** | **−0.714** | **−0.644** | **−0.749** √ | **−0.487** √ | **−0.460** | **−0.479** | **−0.455** |
| irritated | **−0.360** | **−0.307** | **−0.452** √ | **−0.420** | **−0.477** | **−0.313** | **−0.531** √ | **−0.402** |
| happy | **0.544** | **0.487** | **0.622** √ | **0.582** | **0.612** | **0.586** | **0.637** √ | **0.636** |
| afraid | **−0.470** | **−0.548** √ | **−0.449** | **−0.544** | **−0.405** | **−0.393** | **−0.428** √ | **−0.404** |
| relaxed | **0.726** | **0.715** | **0.781** √ | **0.773** | **0.614** √ | **0.466** | **0.608** | **0.549** |
| emphatic | **0.634** | **0.654** √ | **0.569** | **0.592** | **0.676** | **0.690** | **0.681** | **0.723** √ |
| bored | **−0.591** | **−0.530** | **−0.649** √ | **−0.605** | **−0.360** | **−0.381** | **−0.516** √ | **−0.474** |

**Table F.6.** Results of correlation analyses for 4 linguistic measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Level | | | Span | | | | |
| Feature | meanf0 | medianf0 | | ± 2sds mean | 90% Range | | 80% Range | |
| confident | -0.036 | -0.076 | | 0.094 | 0.006 | | 0.091 | |
| tense | 0.043 | 0.110 | | 0.076 | 0.085 | | -0.009 | |
| harsh | 0.234 | 0.265 | | **0.361** | **0.383** | | **0.407** | √ |
| expressive | -0.80 | -0.127 | | 0.068 | 0.005 | | 0.023 | |
| deep | **−0.821** | **−0.842** | √ | **−0.656** | **−0.733** | √ | **−0.463** | |
| weak | 0.074 | 0.127 | | -0.177 | -0.029 | | **−0.324** | √ |
| irritated | 0.134 | 0.207 | | 0.004 | 0.132 | | 0.073 | |
| happy | 0.161 | 0.073 | | 0.145 | 0.105 | | -0.012 | |
| afraid | 0.111 | 0.080 | | -0.018 | -0.078 | | -0.179 | |
| relaxed | -0.239 | -0.295 | | -0.152 | -0.144 | | -0.128 | |
| emphatic | 0.098 | 0.065 | | 0.292 | 0.159 | | 0.240 | |
| bored | 0.183 | 0.196 | | 0.051 | 0.0117 | | -0.027 | |
| | Filtered Speech | | | | | | | |
| | Level | | | Span | | | | |
| | meanf0 | medianf0 | | ± 2sds mean | 90% Range | | 80% Range | |
| confident | -0.144 | -0.253 | | 0.047 | -0.065 | | -0.044 | |
| tense | **0.398** | **0.438** | √ | 0.173 | **0.338** | √ | 0.094 | |
| harsh | -0.058 | -0.082 | | 0.012 | 0.058 | | -0.016 | |
| expressive | 0.042 | -0.036 | | **0.316** | 0.176 | | **0.404** | √ |
| deep | **−0.737** | **−0.730** | √ | **−0.697** | **−0.707** | √ | **−0.531** | |
| weak | **0.361** | **0.399** | √ | 0.221 | **0.347** | √ | 0.174 | |
| irritated | 0.038 | 0.092 | | -0.030 | -0.024 | | -0.082 | |
| happy | 0.026 | -0.054 | | 0.229 | 0.125 | | 0.194 | |
| afraid | **0.450** | **0.486** | √ | 0.249 | **0.389** | √ | 0.187 | |
| relaxed | **−0.476** | **−0.540** | √ | -0.198 | **−0.344** | √ | -0.021 | |
| emphatic | -0.071 | -0.127 | | 0.090 | 0.063 | | 0.177 | |
| bored | -0.183 | -0.147 | | -0.229 | -0.228 | | -0.197 | |

**Table F.7.** Results of correlation analyses for 2 long term distributional measures of level and 4 long term distributional measures of span (measured in Hz) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | |
|---|---|---|---|---|---|---|
| | Level | | | Span | | |
| Feature | meanf0 | medianf0 | | ± 2sds mean | 90% Range | 80% Range |
| confident | -0.068 | -0.125 | | 0.161 | 0.116 | 0.112 |
| tense | 0.291 | **0.369** | √ | 0.001 | 0.083 | -0.064 |
| harsh | **0.414** | **0.437** | √ | 0.222 | 0.278 | 0.244 |
| expressive | -0.213 | -0.274 | | 0.057 | -0.027 | 0.108 |
| deep | **−0.726** | **−0.733** | √ | **−0.679** | **−0.692** | √ **−0.560** |
| weak | 0.237 | **0.317** | √ | -0.032 | 0.062 | -0.113 |
| irritated | 0.066 | 0.148 | | -0.082 | 0.010 | -0.044 |
| happy | 0.014 | -0.087 | | 0.163 | 0.018 | 0.162 |
| afraid | **0.356** | **0.385** | √ | 0.099 | 0.182 | 0.048 |
| relaxed | -0.084 | -0.151 | | 0.229 | 0.104 | 0.295 |
| emphatic | -0.147 | -0.190 | | 0.156 | -0.015 | 0.213 |
| bored | 0.048 | 0.119 | | -0.172 | -0.060 | -0.212 |
| | Filtered Speech | | | | | |
| | Level | | | Span | | |
| | meanf0 | medianf0 | | ± 2sds mean | 90% Range | 80% Range |
| confident | -0.148 | -0.219 | | 0.087 | 0.017 | 0.046 |
| tense | 0.187 | 0.257 | | -0.033 | 0.083 | -0.028 |
| harsh | -0.134 | -0.106 | | -0.209 | -0.200 | -0.232 |
| expressive | -0.007 | -0.118 | | 0.288 | 0.131 | 0.311 |
| deep | **−0.882** | √ **−0.868** | | **−0.809** | √ **−0.790** | **−0.617** |
| weak | 0.044 | 0.122 | | -0.028 | 0.038 | 0.064 |
| irritated | -0.094 | -0.021 | | -0.196 | -0.096 | -0.061 |
| happy | -0.170 | -0.264 | | 0.009 | -0.072 | 0.063 |
| afraid | 0.103 | 0.177 | | 0.015 | 0.068 | 0.149 |
| relaxed | **−0.417** | **−0.486** | √ | -0.191 | **−0.349** | √ -0.094 |
| emphatic | -0.086 | -0.172 | | 0.195 | 0.111 | 0.212 |
| bored | -0.079 | -0.005 | | **−0.338** | √ -0.260 | -0.262 |

**Table F.8.** Results of correlation analyses for 2 long term distributional measures of level and 4 long term distributional measures of span (measured in Hz) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Level | | | | Span | | | | | |
| Feature | meanf0 | | medianf0 | | ± 2sds mean | | 90% Range | | 80% Range | |
| confident | -0.036 | | -0.076 | | 0.094 | | 0.006 | | 0.091 | |
| tense | 0.043 | | 0.110 | | 0.076 | | 0.085 | | -0.009 | |
| harsh | 0.234 | | 0.265 | | **0.361** | | **0.383** | | **0.407** | √ |
| expressive | -0.080 | | -0.127 | | 0.068 | | 0.005 | | 0.023 | |
| deep | **−0.821** | | **−0.842** | √ | **−0.656** | | **−0.733** | √ | **−0.463** | |
| weak | 0.074 | | 0.127 | | -0.177 | | -0.029 | | **−0.324** | √ |
| irritated | 0.134 | | 0.207 | | 0.004 | | 0.132 | | 0.073 | |
| happy | 0.161 | | 0.073 | | 0.145 | | 0.105 | | -0.012 | |
| afraid | 0.111 | | 0.080 | | -0.018 | | -0.078 | | -0.179 | |
| relaxed | -0.239 | | -0.295 | | -0.152 | | -0.144 | | -0.128 | |
| emphatic | 0.098 | | 0.065 | | 0.292 | | 0.159 | | 0.240 | |
| bored | 0.183 | | 0.196 | | 0.051 | | 0.117 | | -0.027 | |
| | Filtered Speech | | | | | | | | | |
| | Level | | | | Span | | | | | |
| | meanf0 | | medianf0 | | ± 2sds mean | | 90% Range | | 80% Range | |
| confident | -0.144 | | -0.253 | | 0.047 | | -0.065 | | -0.044 | |
| tense | **0.398** | | **0.438** | √ | 0.173 | | **0.338** | √ | 0.094 | |
| harsh | -0.058 | | -0.082 | | 0.012 | | 0.058 | | -0.016 | |
| expressive | 0.042 | | -0.036 | | **0.316** | | 0.176 | | **0.404** | √ |
| deep | **−0.737** | √ | **−0.730** | | **−0.697** | | **−0.707** | √ | **−0.531** | |
| weak | **0.361** | | **0.399** | √ | 0.221 | | **0.347** | √ | 0.174 | |
| irritated | 0.038 | | 0.092 | | -0.030 | | -0.024 | | -0.082 | |
| happy | 0.026 | | -0.054 | | 0.229 | | 0.125 | | 0.194 | |
| afraid | **0.450** | | **0.486** | √ | 0.249 | | **0.389** | √ | 0.187 | |
| relaxed | **−0.476** | | **−0.540** | √ | -0.198 | | **−0.344** | √ | -0.021 | |
| emphatic | -0.071 | | -0.127 | | 0.090 | | 0.063 | | 0.177 | |
| bored | -0.183 | | -0.147 | | -0.229 | | -0.288 | | -0.197 | |

**Table F.9.** Results of correlation analyses for 2 long term distributional measures of level and 4 long term distributional measures of span (measured in ERB) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Level | | | Span | | | | | |
| Feature | meanf0 | medianf0 | | ± 2sds mean | | 90% Range | | 80% Range | |
| confident | -0.068 | -0.125 | | 0.161 | | 0.116 | | 0.112 | |
| tense | 0.291 | **0.369** | √ | 0.001 | | 0.083 | | -0.064 | |
| harsh | **0.414** | **0.437** | √ | 0.222 | | 0.278 | | 0.244 | |
| expressive | -0.213 | -0.274 | | 0.057 | | -0.027 | | 0.108 | |
| deep | **−0.726** | **0.733** | √ | **−0.679** | | **−0.692** | √ | **−0.560** | |
| weak | 0.237 | **0.317** | √ | -0.032 | | 0.062 | | -0.113 | |
| irritated | 0.066 | 0.148 | | -0.082 | | 0.010 | | -0.044 | |
| happy | 0.014 | -0.087 | | 0.163 | | 0.018 | | 0.162 | |
| afraid | **0.356** | **0.385** | √ | 0.099 | | 0.182 | | 0.048 | |
| relaxed | -0.084 | -0.157 | | 0.229 | | 0.104 | | 0.295 | |
| emphatic | -0.147 | -0.190 | | 0.156 | | -0.015 | | 0.213 | |
| bored | 0.048 | 0.119 | | -0.172 | | -0.060 | | -0.212 | |
| | Filtered Speech | | | | | | | | |
| | Level | | | Span | | | | | |
| | meanf0 | medianf0 | | ± 2sds mean | | 90% Range | | 80% Range | |
| confident | -0.148 | -0.219 | | 0.087 | | 0.017 | | 0.046 | |
| tense | 0.187 | 0.257 | | -0.033 | | 0.083 | | -0.028 | |
| harsh | -0.134 | -0.106 | | -0.209 | | -0.200 | | -0.232 | |
| expressive | -0.007 | -0.118 | | 0.288 | | 0.131 | | **0.311** | √ |
| deep | **−0.882** | **−0.868** | √ | **−0.809** | √ | **−0.790** | | **−0.617** | |
| weak | 0.044 | 0.122 | | -0.028 | | 0.038 | | 0.064 | |
| irritated | -0.094 | -0.021 | | -0.196 | | -0.096 | | -0.061 | |
| happy | -0.170 | -0.264 | | 0.009 | | -0.072 | | 0.063 | |
| afraid | 0.103 | 0.177 | | 0.015 | | 0.068 | | 0.149 | |
| relaxed | **−0.417** | **−0.486** | √ | -0.191 | | **−0.349** | √ | -0.094 | |
| emphatic | 0.086 | -0.172 | | 0.195 | | -0.111 | | 0.212 | |
| bored | -0.079 | -0.005 | | **−0.388** | √ | -0.260 | | -0.262 | |

**Table F.10.** Results of correlation analyses for 2 long term distributional measures of level and 4 long term distributional measures of span (measured in ERB) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech Span | | | | Filtered Speech Span | | | |
|---|---|---|---|---|---|---|---|---|
| Feature | ± 2sds mean | | 90% Range | 80% Range | ± 2sds mean | | 90% Range | 80% Range |
| confident | **0.333** | √ | 0.004 | 0.089 | **0.426** | √ | -0.163 | 0.008 |
| tense | -0.176 | | 0.054 | -0.222 | **−0.303** | √ | -0.241 | -0.198 |
| harsh | 0.189 | | 0.229 | 0.132 | 0.146 | | 0.002 | 0.028 |
| expressive | **0.414** | √ | 0.076 | 0.087 | **0.617** | √ | 0.187 | **0.424** |
| deep | 0.276 | | -0.290 | 0.170 | 0.073 | | -0.295 | 0.016 |
| weak | **−0.499** | √ | -0.039 | **−0.454** | -0.216 | | 0.248 | -0.142 |
| irritated | **−0.317** | √ | 0.222 | 0.014 | -0.240 | | -0.145 | -0.279 |
| happy | 0.150 | | -0.132 | -0.129 | **0.467** | √ | 0.039 | 0.157 |
| afraid | -0.285 | | **−0.347** | **−0.374** | -0.247 | | 0.232 | -0.061 |
| relaxed | 0.234 | | -0.012 | 0.027 | **0.503** | √ | -0.133 | **0.369** |
| emphatic | **0.428** | √ | -0.014 | 0.150 | **0.460** | √ | 0.134 | **0.347** |
| bored | -0.274 | | -0.014 | -0.232 | **−0.347** | √ | -0.266 | -0.230 |

**Table F.11.** Results of correlation analyses for 4 long term distributional measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the MTV passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

| | Normal Speech Span | | | | Filtered Speech Span | | | |
|---|---|---|---|---|---|---|---|---|
| Feature | ± 2sds mean | | 90% Range | 80% Range | ± 2sds mean | | 90% Range | 80% Range |
| confident | **0.512** | √ | 0.154 | 0.178 | **0.345** | √ | -0.056 | 0.080 |
| tense | **−0.663** | √ | -0.031 | **−0.334** | **−0.536** | √ | 0.078 | -0.124 |
| harsh | -0.188 | | 0.099 | 0.026 | **−0.306** | √ | -0.213 | -0.164 |
| expressive | **0.546** | √ | 0.063 | 0.289 | **0.591** | √ | 0.002 | 0.295 |
| deep | 0.195 | | -0.213 | 0.021 | 0.151 | | -0.276 | 0.050 |
| weak | **−0.551** | √ | -0.083 | **−0.382** | **−0.343** | √ | 0.121 | 0.025 |
| irritated | **−0.388** | √ | -0.083 | -0.100 | **−0.376** | √ | 0.138 | 0.032 |
| happy | **0.462** | √ | -0.100 | 0.221 | **0.512** | √ | -0.113 | 0.215 |
| afraid | **−0.470** | √ | -0.035 | -0.228 | -0.269 | | 0.171 | 0.175 |
| relaxed | **0.716** | √ | 0.156 | **0.445** | **0.424** | √ | **−0.316** | 0.152 |
| emphatic | **0.550** | √ | 0.068 | **0.321** | **0.565** | √ | 0.083 | 0.276 |
| bored | **−0.525** | √ | -0.068 | **−0.304** | **−0.457** | √ | -0.123 | -0.069 |

**Table F.12.** Results of correlation analyses for 4 long term distributional measures of span (measured in semitones) with listener judges' ratings of 12 speaker characteristics reading the Railways passage, for both normal and filtered speech. In this table, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation coefficient that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

# Bibliography


ADDINGTON, D. W. 1968. The relationship of selected vocal characteristics to personality perception. *Speech Monographs* 25.492–503.

ALLPORT, G. W., & H. CANTRAL. 1934. Judging personality from voice. *Journal of Social Psychology* 5.37–54.

ARVANITI, A., D. R. LADD, & I. MENNEN. 1998. Stability of tonal alignment: the case of Greek prenucleur accents. *Journal of Phonetics* 26.3–25.

BECKMAN, M.E., & J.B. PIERREHUMBERT. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3.255–310.

BEZOOIJEN, R. VAN. 1984. *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht, The Netherlands: Foris.

BIRNBAUM, M.H., & C.T. VEIT. 1974. Scale conversion as a criterion for rescaling: Information integration with difference, ratio and averaging tasks. *Perception and Psychophysics* 15.

BOLINGER, D. 1986. *Intonation and its parts: melody in spoken English*. Palo Alto, CA: Stanford University Press.

BRAUN, A., & T. RIETVELD. 1995. The influence of smoking habits on perceived age. In *Proceedings of the XIIIth international congress of phonetic sciences*, volume 2, 294–297, Stockholm.

BROWN, B. L., W. J. STRONG, & A. C. RENCHER. 1973. Perceptions of personality from speech: effects of manipulations of acoustical parameters. *Journal of the Acoustical Society of America* 54.29–35.

——, ——, & ——. 1974. 54 voices from 2: the effects of simultaneous manipulations of rate, fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. *Journal of the Acoustical Society of America* 55.313–318.

BRUCE, G., & E. GÅRDING. 1978. A prosodic typology for Swedish dialects. In *Nordic Prosody*, ed. by E. Gårding, G. Bruce, & R. Bannert, 219–228. Gleerup.

CHEYNE, W. 1970. Stereotyped reactions to speakers with Scottish and English regional accents. *British Journal of Social and Clinical Psychology* 9.77–79.

CLARK, R. A. J. 1999. Using prosodic structure to improve pitch range variation in text to speech synthesis. In *Proceedings for the 14th International Congress of Phonetic Sciences*, San Francisco.

COHEN, A., & J.'T HART. 1967. On the anatomy of intonation. *Lingua* 19.177–92.

CONNELL, B., & D. ROBERT LADD. 1990. Aspects of pitch realisation in Yoruba. *Phonology* 1–30.

COOPER, W. E., & J. M. SORENSEN. 1981. *Fundamental Frequency in Sentence Production*. New York: Springer-Verlag.

COSMIDES, L. 1983. Invariances in the acoustic expression of emotion in speech. *Journal of Experimental Psychology: Human Perception and Performance* 9.864–881.

CRYSTAL, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge University Press.

CURRY, E. T. 1940. The pitch characteristics of the adolescent male voice. *Speech Monographs* 7.48–62.

CUTLER, A., & D. R. LADD (eds.) 1983. *Prosody: Models and Measurements*, chapter 11, 141–146. Springer-Verlag.

DAVITZ, J. R. 1969. *The Language of Emotion*, volume 6 of *Personality and psychopathology*. New York: Academic Press.

DOMMELEN, W. A. VAN, & B. H. MOXNESS. 1995. Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech* 38.267–287.

DORDAIN, M., C. CHEVRIE-MULLER, & F. GRÉMY. 1967. Etude clinique et instrumentale de la voix et de parole des femmes agées. *Revue français de gerontologie* 13.163–170.

EARLE, M. A. 1975. *An acoustic phonetic study of North Vietnamese tones*. Monograph 11. Santa Barbara: Speech Communication Research Laboratories Inc.

FAIRBANKS, G., & W. PRONOVOST. 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs* 6.87–104.

FRØKJAER-JENSEN, B., & S. PRYTZ. 1976. Registration of voice quality. *Bruel and Kjaer Technical Review* 3.3–17.

GILES, H. 1979. Ethnicity markers in speech. In *Social Markers in Speech*, ed. by K. R. Scherer & H. Giles, chapter 7, 251–290. Cambridge University Press.

——, & P. F. POWESLAND. 1975. *Speech Style and Social Evaluation*. European Monographs in Social Psychology. London: Academic Press.

GRADDOL, D. 1986. Discourse specific pitch behaviour. In *Intonation in Discourse*, ed. by C. Johns-Lewis, 221–237. London: Croom Helm.

GÅRDING, E. 1983. A generative model of intonation. In *Prosody: models and measurement*, ed. by A. Cutler & D.R. Ladd, 11–25. Springer-Verlag.

GREASLEY, P., C. SHERRARD, M. WATERMAN, J. SETTER, P. ROACH, S. ARNFIELD, & D. HORTON. 1996. The perception of emotion in speech. *International Journal of Psychology* 31.4763.

HATCH, E., & A. LAZARATON. 1991. *The Research Manual: Design and Statistics for Applied Linguistics*. Newbury House.

HELMHOLTZ, H. L. F. 1954. *On the Sensations of Tone*. New York: Dover.

HENTON, C. G. 1989. Fact and fiction in the description of female and male pitch. *Language and Communication* 9.299–311.

—— 1995. Pitch dynamism in female and male speech. *Language and Communication* 15.43–61.

HERMES, D. J., & H. H. RUMP. 1994. Perception of prominence in speech intonation induced by rising and falling pitch movements. *Journal of the Acoustical Society of America* 96.83–92.

——, & J. C. VAN GESTEL. 1991. The frequency scale of speech intonation. *Journal of the Acoustical Society of America* 90.97–102.

HIRSCHBERG, J., & J. PIERREHUMBERT. 1986. Intonational structuring of discourse. In *Proceedings of the twenty-fourth meeting of the Association for Computational Linguistics*, 136–144, New York.

HOLLIEN, H., D. DEW, & P. PHILIPS. 1971. Phonational frequency ranges of adults. *Journal of Speech and Hearing Research* 14.755–760.

——, & F. T. SHIPP. 1972. Speaking fundamental frequency and chronological age in males. *Journal of Speech and Hearing Research* 15.155–159.

HORII, Y. 1975. Some statistical characteristics of voice fundamental frequency. *Journal of Speech and Hearing Research* 18.192–101.

HUTTAR, G. 1968. Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech and Hearing Research* 11.481–487.

JASSEM, W. 1971. Pitch and compass of the speaking voice. *Journal of the International Phonetic Association* 1.59–68.

—— 1975. Normalisation of f0 curves. In *Auditory Analysis and Perception of Speech*, ed. by G. Fant & M. Tatham, 523–530. London: Academic Press.

KRAAYEVELD, H., 1997. *Idiosyncrasy in Prosody: Speaker and speaker group identification in Dutch using melodic and temporal information*. Catholic University, Nijmegen dissertation.

LADD, D. R. 1996. *Intonational Phonology*. Cambridge University Press.

——, & A. CUTLER. 1983. Models and measurements in the study of prosody. In *Prosody: models and measurement*, ed. by A. Cutler & D. R. Ladd, 1–10. Springer-Verlag.

——, K. E. A. SILVERMAN, F. TOLKMITT, G. BERGMANN, & K. R. SCHERER. 1985. Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *Journal of the Acoustical Society of America* 78.435–444.

——, & J. TERKEN. 1995. Modelling intras- and inter-speaker pitch range variation. In *Proceedings of the International Conference of Phonetic Sciences*, volume 2, 386–389, Stockholm.

LAMBERT, W. 1972a. Evaluational reactions to spoken languages. In *Language, Psychology and Culture*, ed. by A. S. Dil, 80–96. Palo Alto: Stanford University Press.

—— 1972b. A social psychology of bilinguilism. In *Language, Psychology and Culture*, ed. by A. S. Dil, 212–235. Palo Alto: Stanford University Press.

LASS, N. J., A. S. BEVERLY, D. K. NICOSIA, & L. A. SIMPSON. 1978. An investigation of speaker height and weight identification by means of direct estimation. *Journal of Phonetics* 6.69–76.

LAVER, J., & R. HANSON. 1981. Describing the normal voice. In *Speech Evaluation in Psychiatry*, ed. by J. K. Darby, chapter 3, 51–78. Grune and Statton.

——, & P. TRUDGILL. 1979. Phonetic and linguistic markers in speech. In *Social Markers in Speech*, ed. by K. R. Scherer & H. Giles, chapter 1, 1–32. Cambridge University Press.

LEINONEN, L., T. HILTUNEN, L. LINNANKOSKI, & M. L. LAAKSO. 1997. Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America* 102.922–927.

LIBERMAN, M., & J. PIERREHUMBERT. 1984. Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, ed. by M. Aronoff & R. T. Oehrle, chapter 10, 157–233. MIT Press.

LIEBERMAN, P., & S. B. MICHAELS. 1962. Some aspects of fundamental fequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America* 34.922–927.

LODGE, M. 1981. *Magnitude scaling: Quantitative measurement of opinions*. Number 25 in Quantitative applications in the social sciences. London: Sage.

MAEDA, S., 1976. *A characterization of American English Intonation*. MIT dissertation.

MARKEL, J. D., B. T. OSHIKA, & A. H. GRAY. 1977. Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 25.330–337.

MEEK, PAULA M., LEE SENNOTT-MILLER, & SANDRA L. FERKETICH. 1992. Scaling stimuli with magnitude estimation. *Research in Nursing and Health* 15.77–81.

MELLERS, B. A. 1983. Evidence against "absolute" scaling. *Perception and Psychophysics* 33.523–526.

MENN, L., & S. BOYCE. 1982. Fundamental frequency and discourse structure. *Language and Speech* 25.341–383.

MONAGHAN, A. I. C., & D.R . LADD. 1990. Speaker-dependent and speaker-independent parameters in intonation. In *Proceedings of the ESCA Workshop on Speaker Characterisation*, Edinburgh.

MOORE, B. C. J. 1997. *An Introduction to the Psychology of Hearing*. Academic Press, fourth edition.

MOZZICONACCI, S., 1998. *Speech Variability and Emotion: Production and Perception*. University of Eindhoven dissertation.

MURRAY, I. R., & J. L. ARNOTT. 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* 93.1097–1108.

OSTENDORF, M., P. J. PRICE, & S. SHATTUCK-HUFNAGEL. 1995. The Boston University radio news corpus. Technical Report ECS-95-001, Boston University, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.

PAKOSZ, M. 1982. Intonation and attitude. *Lingua* 56.153–178.

PATTERSON, R. D. 1976. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America* 59.

PIERREHUMBERT, J. B., 1980. *The Phonology and Phonetics of English Intonation*. Cambridge MA: MIT dissertation.

——, & M. E. BECKMAN. 1988. *Japanese Tone Structure*. Cambridge, M.A.: MIT Press.

PTACEK, P. H., E. K. SANDER, W. H. MALONEY, & C. C. ROE-JACKSON. 1966. Phonatory and related changes with advanced age. *Journal of Speech and Hearing Research* 9.175–184.

RAMIG, L. O., & R. L. RINGEL. 1983. Effects of physiological aging on selected acoustic charactersitics of voice. *Journal of Speech and Hearing* 26.22–30.

RITSMA, J. R. 1967. Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America* 42.

ROSE, P., 1982. *An acoustically based phonetic description of the syllable in the Zhenhai dialect*. Cambridge University dissertation.

—— 1987. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication* 6.343–351.

—— 1991. How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication* 10.229–247.

SCHERER, K. R. 1979. Personality markers in speech. In *Social Markers in Speech*, chapter 5, 147–210. Cambridge University Press.

—— 1981. Speech and emotional states. In *Speech Evaluation in Psychiatry*, chapter 10, 189–220. Grune and Stratton.

—— 1986. Vocal affect expression: a review and a model for future research. *Psychological Bulletin* 99.143–165.

—— 1988. *Facets of emotion; recent research*. Hillsdale, NJ: Erlbaum.

——, & P. EKMAN (eds.) 1982. *Handbook of Methods in Nonverbal Behaviour Research*. Cambridge University Press.

——, D. R. LADD, & K. E. A. SILVERMAN. 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America* 76.1346–1356.

SENNOT-MILLER, L., C. MURDAUGH, & A. S. HINSHAW. 1988. Magnitude estimation: Issues and practical application. *Western Journal of Nursing Research* 10.

SHRIBERG, E., D. R. LADD, J. TERKEN, & A. STOLCKE. 1996. Modeling pitch range variation within and across speakers: predicting fo targets when "speaking up". In *Proceedings for the International Conference on Spoken Language Processing*, 1–4, Philadelphia, PA, USA. Addendum.

SILVERMAN, K. 1986. F0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica* 43.76–92.

——, 1987. *The structure and processing of fundamental frequency contours.* Cambridge University dissertation.

——, M. BECKMAN, J. PITRELLI, M. OSTENDORF, C. WIGHTMAN, P. PRICE, J. PIERREHUMBERT, & J. HIRSCHBERG. 1992. Tobi: A standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing*, 867–870.

STEVENS, S. S. 1957. On the psychophysical law. *Psychological Review* 64.153–181.

—— 1974. Measurement. In *Scaling: A Sourcebook for Behavioural Scientists*, ed. by G. M. Maranell, chapter 2, 22–41. Chicago: Aldine.

——, J. VOLKMANN, & E. B. NEWMAN. 1937. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America* 8.185–190.

'T HART, J., R. COLLIER, & A. COHEN. 1990. *A Perceptual Study of Intonation*. Cambridge University Press.

TAKEFUTA, Y. 1975. Method of acoustic analysis of intonation. In *Measurement Procedures in Speech Hearing and Language*, ed. by S. Singh, 363–378. Baltimore: University Park Press.

THORSEN, N. 1978. An acoustical analysis of Danish intonation. *Journal of Phonetics* 6.151–175.

TRAUNMÜLLER, H., & A. ERIKSSON. 1995. The perceptual evaluation of f0 excursions in speech as evidenced in liveliness estimations. *Journal of the Acoustical Society of America* 97.1905–1915.

ULDALL, E. 1960. Attitudinal meanings conveyed by intonation. *Language and Speech* 3.223–34.

—— 1964. Dimensions of meaning in intonation. In *In honour of Daniel Jones*, ed. by D. Abercrombie. London: Longmans.

VAISSIÈRE, J. 1983. Language-independent prosodic features. In *Prosody: models and measurement*, ed. by A. Cutler & D.R. Ladd, 53–66. Springer-Verlag.

WILLIAMS, C. E., & K. N. STEVENS. 1972. Emotions and speech: some acoustic correlates. *Journal of the Acoustical Society of America* 52.1238–1250.

ZEMLIN, W. R. 1981. *Speech and hearing science: anatomy and physiology*. New Jersey: Englewoods Cliffs: Prentice-Hall, 2nd edition.

ZWICKER, E. 1961. Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America* 33.248.